NUS
National University of Singapore

Audio-Visual speech separation: Datasets

Pan Zexu

NUS Graduate School for Integrative Sciences and Engineering

Outlines

1. Audio-visual (AV) speech separation / speaker extraction datasets.
   1.1. Overlapped speech simulation procedure.
   1.2. Grid datasets [1]
   1.3. LRS2 datasets [2]
   1.4. AVSpeech datasets [3]
2. Other AV datasets related to speech front-end processing.
   2.1. AV Active speaker detection: AVA-ActiveSpeaker [4]
   2.2. AV diarization: VoxConverse [5]

[1] M. Cooke, J. Barker, S. Cunningham, and X. Shao. An audio-visual corpus for speech perception and automatic speech recognition. The Journal of the Acoustical Society of America, 120(5):2421–2424, 2006.

[2] Afouras, Triantafyllos, et al. "Deep audio-visual speech recognition." *IEEE transactions on pattern analysis and machine intelligence* (2018).

[3] Ephrat, Ariel, et al. "Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation." *ACM Transactions on Graphics (TOG)* 37.4 (2018): 1-11.

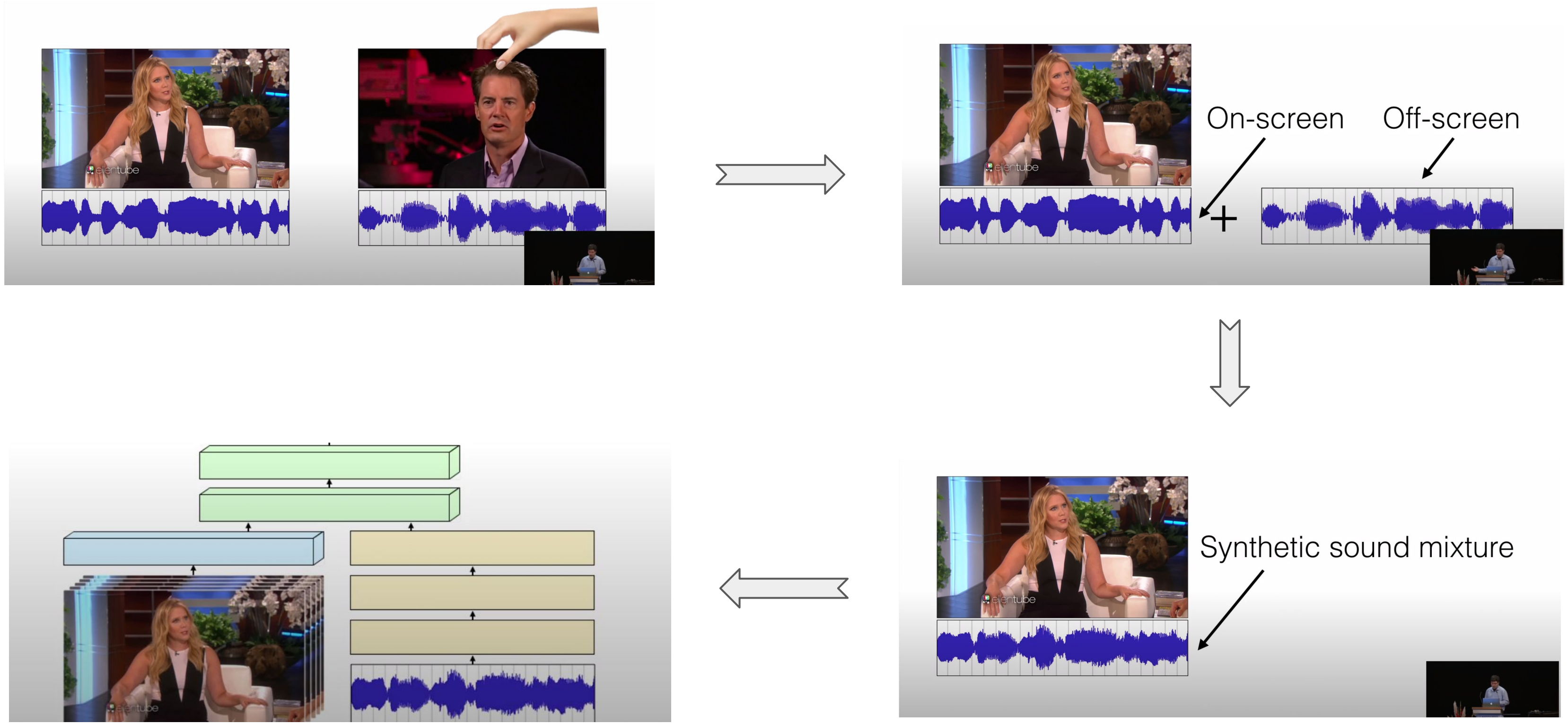
[4] Roth, Joseph, et al. "Ava Active Speaker: An Audio-Visual Dataset for Active Speaker Detection." *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020.

[5] Chung, Joon Son, et al. "Spot the conversation: speaker diarisation in the wild." *arXiv preprint arXiv:2007.01216* (2020).

Outlines

1. Audio-visual (AV) speech separation / speaker extraction datasets.
   1.1. Overlapped speech simulation procedure.
   1.2. Grid dataset [1]
   1.3. LRS2 dataset [2]
   1.4. AVSpeech dataset [3]
2. Other AV datasets related to speech front-end processing.
   2.1. AV Active speaker detection: AVA-ActiveSpeaker [4]
   2.2. AV diarization: VoxConverse [5]

1.1 Overlapped speech simulation



1.1 Datasets used since 2017

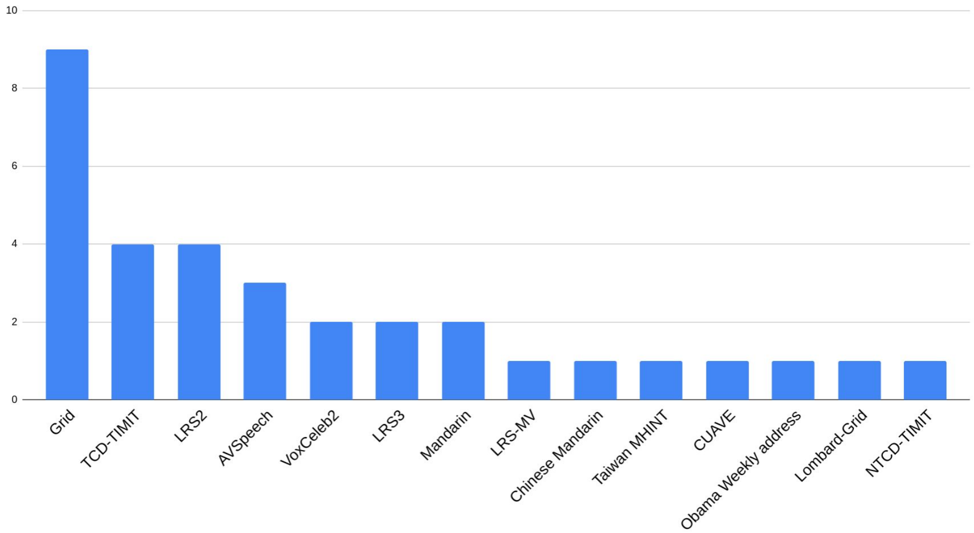

Datasets of 21 AV speech separation paper

Table 1: Statistics of commonly used AV speaker extraction datasets

	Datasets	No. hours	Language	Speakers
Small	Grid	27.5	English	33
	TCD-TIMIT	10	English	62
Medium	LRS2 (BBC)	224	English	-
	LRS3 (TED)	475	English	5543
Large	AVSpeech	4700	Mix	-
	VoxCeleb2	2000	Mix	6112

1.2 Grid dataset

1. Intro
   a. Multitalker audiovisual sentence corpus of 1000 sentences spoken by each of 34 English talkers (18 male, 16 female). Sentences are of the form “put red at G9 now”
   b. The dataset is recorded in lab environment, thus it is background noise free. However, The vocabulary size is very small (51).
   c. Primarily used for audio visual automatic speech recognition. Overlapped speech is simulated if it is used for speech separation.
2. Content
   a. Audio : 3 seconds @ 25 kHz
   b. Video: 3 seconds (750*576 pixels, ~6kbit/s)
   c. Text: transcriptions (eg: place red at C9 again)

1.2 Grid dataset

3. Results on Speech separation
   a. Each paper is using their own simulated datasets. No standard benchmark.
   b. Separate target speech from overlapped speech and background nose: [1], [2]
   c. Separate target speech from overlapped speech: [3],[4],[5],[6],[7]
   d. Speaker dependent: [1],[4]

Table 2 : Results of Audio-visual speech extraction on Grid dataset.

	SNRi	PESQi
Visual Speech enhancement [1]	5.59	0.95
A visual pilot [2]	8.78	0.89
Listen and look [3]	8.64	-
Seeing through noise [4]	5.58	0.5
Audio-Visual deep clustering [5]	8.95	-
Face landmark based [6]	7.84	0.76
Event Driven cameras [7]	6.82	0.71

[1] Gabbay, Aviv, Asaph Shamir, and Shmuel Peleg. "Visual Speech Enhancement." *Proc. Interspeech 2018* (2018): 1170-1174.

[2] Li, Yun, et al. "A Visual-Pilot Deep Fusion for Target Speech Separation in Multitalker Noisy Environment." *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020.

[3] Lu, Rui, Zhiyao Duan, and Changshui Zhang. "Listen and look: Audio–visual matching assisted speech source separation." *IEEE Signal Processing Letters* 25.9 (2018): 1315-1319.

[4] Gabbay, Aviv, et al. "Seeing through noise: Visually driven speaker separation and enhancement." *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018.

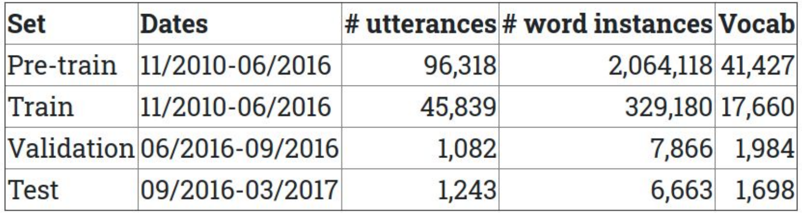
[5] Lu, Rui, Zhiyao Duan, and Changshui Zhang. "Audio–visual deep clustering for speech separation." *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 27.11 (2019): 1697-1712.

[7] Morrone, Giovanni, et al. "Face landmark-based speaker-independent audio-visual speech enhancement in multi-talker environments." *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019.

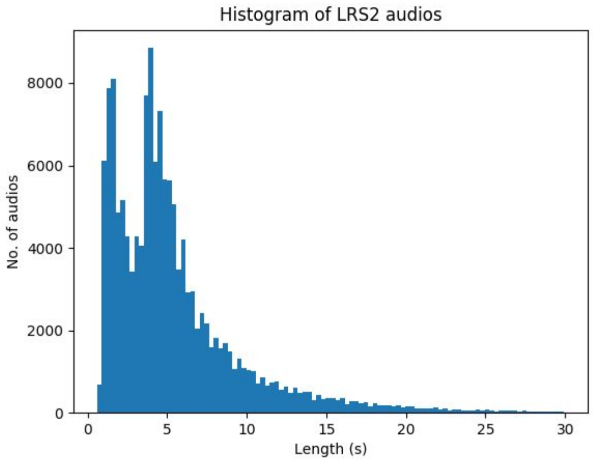
[8] Arriandiaga, Ander, et al. "Audio-Visual Target Speaker Extraction on Multi-Talker Environment using Event-Driven Cameras." *arXiv preprint arXiv:1912.02671* (2019).

1.3 LRS2 dataset

1. Intro
   a. Oxford-BBC Lip reading sentences 2 datasets contains English videos from BBC television. The train, val and test divided according to broadcast date.
   b. It was primarily used for audio-visual automatic speech recognition. Overlapped speech is simulated if it is used for speech separation.
2. Content
   a. Video: 160*160 pixels at 25fps
   b. Audio: Extracted from video with 16kHz.
   c. Text: Transcriptions.



Set	Dates	# utterances	# word instances	Vocab
Pre-train	11/2010-06/2016	96,318	2,064,118	41,427
Train	11/2010-06/2016	45,839	329,180	17,660
Validation	06/2016-09/2016	1,082	7,866	1,984
Test	09/2016-03/2017	1,243	6,663	1,698



1.3 LRS2 dataset

3. Results on Speech Separation
   a. Each paper is using their own simulated datasets. No standard benchmark.

Table 3: Results of Audio-visual speech extraction on LRS2 dataset with two speaker overlapped.

LRS2	SNRi (SNR)	PESQi (PESQ)
Deep audio-visual [1]	(12.6)	(3.18)
The conversation [2]	12.1	1.35
My lips are concealed [3]	12.8	-
Time domain av [4]	(13.03)	-

[1] Li, Chenda, and Yanmin Qian. "Deep Audio-Visual Speech Separation with Attention Mechanism." *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020.
[2] Afouras, Triantafyllos, Joon Son Chung, and Andrew Zisserman. "The Conversation: Deep Audio-Visual Speech Enhancement." *Proc. Interspeech 2018* (2018): 3244-3248.
[3] Afouras, Triantafyllos, Joon Son Chung, and Andrew Zisserman. "My Lips Are Concealed: Audio-Visual Speech Enhancement Through Obstructions}}." *Proc. Interspeech 2019* (2019): 4295-4299.
[4] Wu, Jian, et al. "Time domain audio visual speech separation." *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2019.

1.4 AVSpeech dataset

1. Intro
    a. Audio-visual datasets of youtube video clips with no interfering background noise.
    b. Spanning a variety of people, language and face poses.
    c. Primarily used for audio visual speech enhancement and separation. Overlapped speech is simulated for speech separation.
2. Content
    a. Video: 290k video segments, 3-10 seconds each
3. Results on speech separation
    a. [1] used the full dataset
    b. [2] used a subset of 100 hours

Table 4: Results of Audio-visual speech extraction on AVSpeech dataset with two speaker overlapped.

AVSpeech	SNRi	PESQi
Looking to listen [1]	9.9	-
Video Conferencing [2]	11.34	0.45

[1] Ephrat, Ariel, et al. "Looking to listen at the cocktail party: a speaker-independent audio-visual model for speech separation." *ACM Transactions on Graphics (TOG)* 37.4 (2018): 1-11.

[2] İnan, Berkay, et al. "Evaluating Audiovisual Source Separation in the Context of Video Conferencing}}." *Proc. Interspeech 2019* (2019): 4579-4583.

Outlines

1. Audio-visual (AV) speech separation / speaker extraction datasets.
    1.1. Overlapped speech simulation procedure.
    1.2. Grid dataset
    1.3. LRS2 dataset
    1.4. AVSpeech dataset
2. Other AV datasets related to speech front-end processing.
    2.1. AV Active speaker detection: AVA-ActiveSpeaker
    2.2. AV diarization: VoxConverse

2.1 AVA-ActiveSpeaker

1. Intro:
   a. Active speaker detection: To detect the speaking face in the video given the audio and video.
   b. The dataset consist of 3.65 frames with 38.5 hours of face tracks from TV shows / movies.
   c. Spanning a variety of people, language and face poses.
2. Content:
   a. Video: 153 TV shows/ movies
   b. Label: Bounding box of all face tracks in one frame with label (0: speaking, 1:non_speaking)
   c. Each face track has a (local) identity. A new identity is assigned to the same person if the face tracking algorithm stopped and re-track the same person.

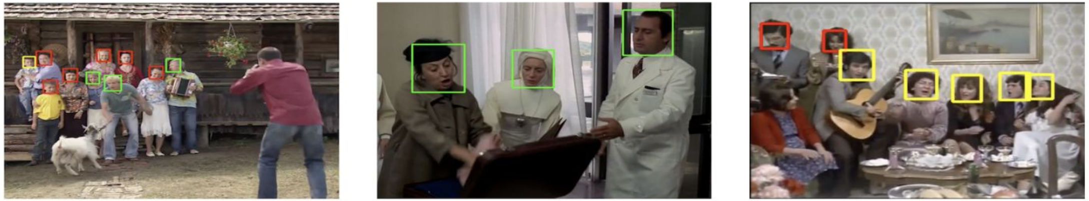


Figure 13: Visualizations of overlapping speaker instances. A green bounding box represents speaking and audible, yellow represents speaking and inaudible, red represents not speaking.

2.2 VoxConverse

1. Intro:
   a. Diarisation: who spoke when in the video. Different from the active speaker detection, we also need to cluster the speakers with their identity globally.
   b. Audio-visual diarisation dataset consist of over 50 hours of multispeaker clips of human speech from youtube videos.
2. Content:
   a. Video: To be confirmed.
   b. Audio:
   c. Label: Rich Transcription Time Marked (RTTM) files
   d. Only development audio files are available now. Visual files are not released yet.
   e. Test set will be released in October 2020 after the VoxCeleb Speaker Recognition Challenge

Conclusion

1. Audio-visual (AV) speech separation / speaker extraction datasets.
   1.1. Overlapped speech simulation procedure.
   1.2. Grid dataset
   1.3. LRS2 dataset
   1.4. AVSpeech dataset
2. Other AV datasets related to speech front-end processing.
   2.1. AV Active speaker detection: AVA-ActiveSpeake
   2.2. AV diarization: VoxConverse

Thank You!