

CineSRD: Leveraging Visual, Acoustic, and Linguistic Cues for Open-World Visual Media Speaker Diarization

Supplementary Material

1. Preprocessing

1.1. Face Detection

The visual input for the active speaker detection model TalkNet is face images. Therefore, we first need to perform face detection on the frames of the video segment corresponding to the line $s \in \mathcal{P}$. We use the RetinaFace [2] model for face detection on raw video frames. RetinaFace efficiently detects faces in images and returns the bounding box, confidence score, and facial keypoints information. Based on this information, we perform subsequent steps such as face cropping and tracking.

1.2. Face Tracking

Face tracking aims to track the same face across consecutive video frames, forming a facial trajectory. We use a simple tracking algorithm based on Intersection over Union (IoU). Specifically, for each detected face in a frame, we attempt to match it with the face from the previous frame. If the IoU of the face bounding boxes from the two frames exceeds a certain threshold (e.g., 0.5), we consider them to be the same face and add it to the current trajectory.

1.3. Face Alignment

Unlike simple video scenes such as meetings or interviews, in visual programs of various genres, the detected faces may exhibit multiple poses due to body positions such as lying flat or reclining. To facilitate subsequent facial feature extraction, we employ an affine transformation method based on facial keypoints for face alignment. Specifically, we use the facial keypoint information returned by RetinaFace in Appendix 1.1, and based on the geometric relationship between these keypoints and predefined standard facial keypoints, we calculate an affine transformation matrix. This matrix is then used to rotate and scale the face image to align it with the standard pose. Figure 1 shows a demonstration of face alignment.

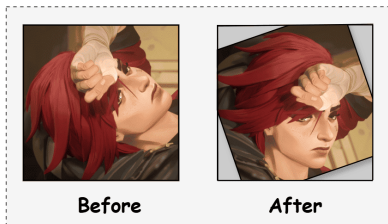


Figure 1. Face alignment.

1.4. Face Quality Evaluation

Face quality assessment is a critical step to ensure the accuracy and reliability of facial feature extraction. We employ the FQA face quality assessment model from Alibaba DAMO Academy, which comprehensively evaluates multiple quality dimensions of a face image, such as sharpness, lighting conditions, and occlusion. Given a face image as input, the model outputs a quality score that reflects the usability of the image. In practice, we assign a quality score to each detected face image within the video of every line and select only the highest-scoring face as the representative face for that line. Figure 2 illustrates a demonstration of face quality evaluation.

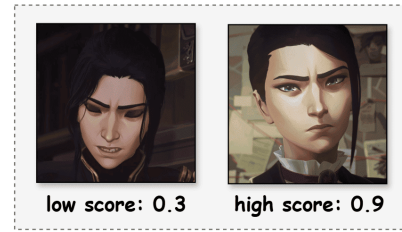


Figure 2. Face quality evaluation.

1.5. Audio Preprocessing

We use moviepy library to extract audio from video files, and then apply Demucs tool [1, 3] for vocal track separation. Isolating human voices from the audio is necessary to prevent environmental noise in films and TV series from interfering with timbre feature extraction.

2. Supplementary Information

2.1. SubtitleSD Demonstration

In this study, we conduct experiments on the SubtitleSD dataset, which is collected from audiovisual programs on the online video platform Anonymous. The dataset contains subtitles and videos of both Chinese and English programs. We manually annotate the speaker (character) for each line in the subtitles to evaluate the performance of our method and the baselines. Table 1 presents examples of English subtitles in the SubtitleSD benchmark, where the "Start" and "End" columns indicate the temporal boundaries of each line in the video, and the "Speaker" column denotes the manually annotated character.

Table 1. Examples of subtitles of SubtitleSD dataset.

Start	End	Text	Speaker
0:17:33.25	0:17:34.50	Please, let me speak!	Jayce's mother
0:17:39.54	0:17:42.83	As a lower house, my voice doesn't carry much weight here.	Jayce's mother
0:17:42.91	0:17:46.45	But as a mother, I have a voice that matters deeply.	Jayce's mother
0:17:47.50	0:17:50.25	My son isn't in his right mind.	Jayce's mother
0:17:51.04	0:17:53.95	A crime like this can't be overlooked. The boy must be punished.	Councilor Allira Salo
0:17:54.91	0:17:59.00	His entire life, he's chased an impossible dream.	Jayce's mother
0:17:59.83	0:18:04.20	What he did was, uh, foolish and unwise.	Jayce's mother
0:18:04.29	0:18:08.66	But he has a good heart. Please, let him come home.	Jayce's mother
0:18:08.75	0:18:12.16	A violation of the Ethos calls for banishment,	Heimerdinger
0:18:12.25	0:18:16.25	but I can sympathize with a young man's dream to change the world.	Heimerdinger
0:18:16.83	0:18:20.58	Perhaps in this matter, a lesser sentence may suffice.	Heimerdinger

Table 2. Prompt of en \Rightarrow zh subtitle translation accuracy evaluation.

Preamble	<p>Given multiple lines of dialog along with corresponding audio segments, your task is to determine whether two adjacent lines are spoken by the same character based on both the textual context/semantics and the audio information.</p> <p>>>>> Task Requirements >>>></p> <ol style="list-style-type: none"> 1. For each line of dialog (with its audio), output <0> or <1> to indicate whether this line and the previous line were spoken by the same character. <0> means different speakers, <1> means the same speaker. 2. Make your judgment by combining the textual context, semantic information, and the audio characteristics of each line. 3. The output results must strictly correspond to the input line order. Do not merge, modify, or reorder any dialog lines. Keep the original text and output format exactly as specified.
Format	<p>>>>> Input Format >>>></p> <pre>line1<audio.tag> line2<audio.tag> line3<audio.tag> ...</pre> <p>Judgment results=</p> <p>>>>> Output Format >>>></p> <pre>line1<None> line2<0/1> line3<0/1> ...</pre>
Ending	<p>>>>> Begin the judgment. Follow all the above requirements and format. >>>></p> <pre>{x}</pre> <p>Judgment results=</p>

067 2.2. ALM Prompt

068 We present in Table 2 the prompt format used in the
 069 speaker turn detection module for the ALM to determine
 070 whether two adjacent lines belong to the same speaker. The
 071 prompt is structured as follows:

- 072 1. *Preamble* – an introduction and instructions describing
 073 the task at hand

2. *Format* – specification of the input and output format
3. *Ending* – ending text to prompt the ALM

074
075

076

References

077

078

079

080

081

082

083

084

085

086

087

- [1] Alexandre Défossez. Hybrid spectrogram and waveform source separation. In *Proceedings of the ISMIR 2021 Workshop on Music Source Separation*, 2021. [1](#)
- [2] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5203–5212, 2020. [1](#)
- [3] Simon Rouard, Francisco Massa, and Alexandre Défossez. Hybrid transformers for music source separation. In *ICASSP 23*, 2023. [1](#)