Character-aware audio-visual subtitling in context

Jaesung Huh and Andrew Zisserman

Visual Geometry Group, Department of Engineering Science, University of Oxford
{jaesung,az}@robots.ox.ac.uk

Abstract. This paper presents an improved framework for character-aware audio-visual subtitling in TV shows. Our approach integrates speech recognition, speaker diarisation, and character recognition, utilising both audio and visual cues. This holistic solution addresses what is said, when it's said, and who is speaking, providing a more comprehensive and accurate character-aware subtitling for TV shows. Our approach brings improvements on two fronts: first, we show that audio-visual synchronisation can be used to pick out the talking face amongst others present in a video clip, and assign an identity to the corresponding speech segment. This audio-visual approach improves recognition accuracy and yield over current methods. Second, we show that the speaker of short segments can be determined by using the temporal context of the dialogue within a scene. We propose an approach using local voice embeddings of the audio, and large language model reasoning on the text transcription. This overcomes a limitation of existing methods that they are unable to accurately assign speakers to short temporal segments. We validate the method on a dataset with 12 TV shows, demonstrating superior performance in speaker diarisation and character recognition accuracy compared to existing approaches. Project page : https://www.robots.ox.ac.uk/~vgg/research/llr-context/

Keywords: Character-aware audio-visual subtitling · Audio-visual learning · Video understanding

1 Introduction

Character-aware audio-visual subtitling is an emerging area that aims to automatically generate subtitles for TV shows and movies, including the corresponding speaker names. This task involves determining three key aspects: *what* is being said, *when* it is said, and *who* is saying it. This capability is essential for the audio-impaired, so that they can follow video material – indeed it is a requirement of Subtitles for Deaf and Hard-of-hearing (SDH [69]) that the subtitles include information about speaker identification, as well as information on sound effects and music. It also enables the annotation of large-scale video datasets for training the next generation of visual-language models, capable of learning a higher-level story understanding of video material.

The task builds on developments in three specialised areas: *Automatic speech recognition* (ASR, or speech-to-text) that is primarily concerned with transcribing spoken words into text – determining 'what is spoken'; *Speaker diarisation*,

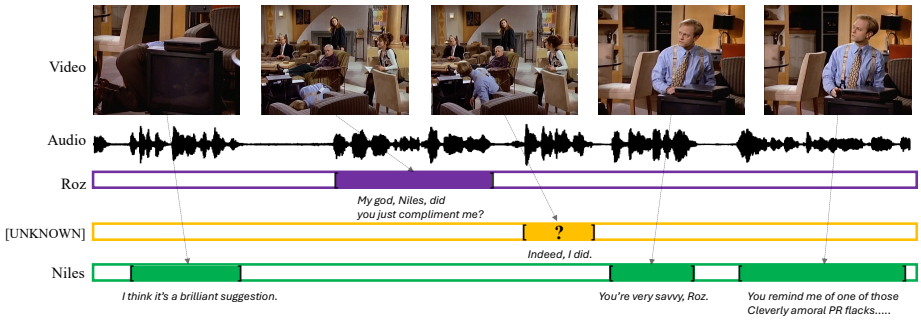

Fig. 1: An example video clip and output of our method. Dialogues in TV shows typically flow continuously, and speaker identities can often be inferred from the content and context of the conversation. In some cases, it's possible to diarise speakers solely based on textual context. Even though we cannot see the speaker visually – so have no evidence from lip-movement – we can infer that the utterance with a question mark (**?**) belongs to 'Niles' by looking at temporal context of the dialogue.

that aims to organise multi-speaker audio into homogeneous single speaker segments, effectively solving 'who spoke when'; and *Character recognition*, that aims to identify the characters appearing in the video clips. Each of these areas is well explored, and can use single modality methods (*i.e.* audio only or visual only) or audio-visual methods. For example, ASR can be audio-only [32,37,62], or audio-visual [2,29,53,65]. Similarly, common methods can be used across the areas. For example, voice embeddings can be used for diarisation by clustering [42,74,76], and for recognition by matching to a gallery of voices [41,49,68]. However, because these are somewhat independent areas, they do not alone provide all the ingredients required.

Recent works have introduced methods and datasets for character-aware audio-visual subtitling, building on elements from the three areas above [44,48]. The state-of-the-art method of Korbar *et al.* [44], proceeds in two stages: it first builds a gallery of voice embeddings for each character using audio-visual methods, and then generates the character-aware subtitles using only audio recognition. Despite its accomplishments, however, this method has two significant shortcomings when determining the character speaking during a temporal segment: (1) it has a poor performance for short segments (those lasting less than 2 seconds), often assigning the wrong character; and (2) it has a low yield over all segments, as often it is unable to classify the character.

In this paper we make three contributions. First, we introduce a new method for identifying the speaker for short segments, building on the insight that assignments that cannot be resolved using only local (temporal) information, can often be disambiguated using the *temporal context* of the surrounding dialogue. We investigate two complementary approaches for this task: (i) using speaker recognition, we note that a short utterance in a dialogue may well be spoken by a character with a longer utterance (where the identity is not ambiguous) else-

where in the scene, and the short segment can then be assigned using *local* voice embeddings, rather than voice embeddings from a gallery where audio conditions may substantially differ; (ii) using a large language model (LLM), the identity of the character speaking the short segment can be resolved based on the *content* of the dialogue, as illustrated by the example in Figure 1. The second contribution is to use a *local visual embedding* around the lip motion synchronised with the speech to determine the speaker. This overcomes a limitation of [44], where a CLIP descriptor of the entire frame is used to predict the speaker identity. The use of a local visual embedding leads to higher yield of assignments for speaker segments. Taken together these two contributions significantly improve the performance over that of [44]. As our third contribution, we validate our method on a large evaluation dataset covering 12 TV series. This dataset incorporates the existing dataset used by [44] and additional shows from [48], demonstrating the generalisation ability of our method.

2 Related Work

Several subtasks within this field have already been explored by researchers. *Speech recognition* [18, 32, 55, 62], or speech-to-text, is primarily concerned with transcribing spoken words into text. However, this subtask typically overlooks the timing of speech and fails to identify the speaker. *Speaker diarisation* [22, 25, 59, 74] aims to identify speech regions and assign speaker labels to each person in an audio file. This task clusters speech segments by speaker without necessarily matching them to specific known individuals. *Character recognition* [11, 38, 40, 61, 63], a well-studied topic in computer vision and speech processing, assigns names to characters appearing in scenes. Character-aware audio-visual subtitling requires the integration of all three tasks, utilising both audio and visual cues from the video.

Character recognition in videos. Recognising characters in video [26, 27, 35, 57] is a challenging task due to the presence of multiple characters in a single frame, occlusions, and variations in appearance. Several methods have been proposed to incorporate additional modalities, such as audio [16, 57], or transcripts [12, 26, 27, 35] which are often unavailable. There are a line of works which use speaker diarisation in TV shows and use the result to cluster the speaker identities [13, 64]. However, they simply cluster the speaker identities, not assigning the actual character's name. Our task involves assigning the specific names of speakers in TV shows using a castlist.

Audio-visual speech processing. Numerous studies have examined human conversation from a broad perspective. Given that these interactions primarily occur through speech, a wide range of research focuses on audio-only approaches to various tasks, including speech recognition [8, 32, 62], speaker identification [21, 24, 43], and speaker diarisation [14, 28]. With a rise of the multimodal learning, researchers have started to incorporate visual information such as lip movements

in addition to audio information to improve the performance of these tasks. For example, they use lip movements [2, 65] or faces [17, 22] in addition to audio to improve the performance of this task.

LLM for video understanding. Large Language Models (LLMs) [1, 6, 39, 70] have driven the great progress not only in Natural Language processing but also in computer vision [7, 9, 50, 52] and audio processing [30, 31]. Over the past few years, there has been a plethora of works which leverage LLMs in various video understanding tasks. There are two different approaches to this. The first approach integrates a pretrained LLM with visual or audio backbones as part of the entire model, fine-tuning it to understand multimodal content [33, 54, 67, 77]. The second approach uses LLM separately from video models to improve performance on video understanding tasks [19, 73].

Human conversation datasets. There has been growing interest in audio-visual datasets with rich transcriptions of spoken conversations, including speech transcripts, timestamps and speaker identities. Several datasets exist with annotations for either one of these aspects. LRS series [2, 3] have advanced audio-visual speech recognition technology, but their single-speaker focus limits development of multi-speaker systems for conversational settings. There exist audio-visual speaker diarisation datasets [22, 75] with multiple speakers but do not have a speech transcripts. The AMI-Corpus [46] and VoxMM [47] are multimodal datasets which provide audio-visual data with speaker identities and speech transcripts. However, both focuses on different domain than ours such as meeting scenarios, commercial or interviews. *Bazinga!* [48] offers rich transcriptions of TV shows, including word-level timestamps, speech transcripts, and speaker identities. We use this dataset to verify our pipeline's performance.

Relation to the method of Korbar et al. [44]. This work also aims to generate character-aware subtitle generation. However, it has several limitations. Firstly, it fails to utilise spatial information from lip-moving areas, which could significantly enhance speaker recognition accuracy. The method utilises CLIP-PAD [45] which recognises characters in scenes without employing a face detection model to identify clips for single-speaker regions. Unfortunately, these clips may contain multiple faces, potentially confusing the model when tasked with identifying the actual speaker. Secondly, it doesn't take advantage of the time-based context when matching speaker names to parts of speech. In TV shows, conversations usually progress in a continuous manner. As a result, it's often possible to figure out who is speaking by considering the overall flow of the dialogue and the context in which things are said.

3 Assigning Speakers to Short Audio Segments

The task here is to assign speaker identities to short temporal speech segments. We assume that we have a gallery/library of voice embeddings available for the principal characters.

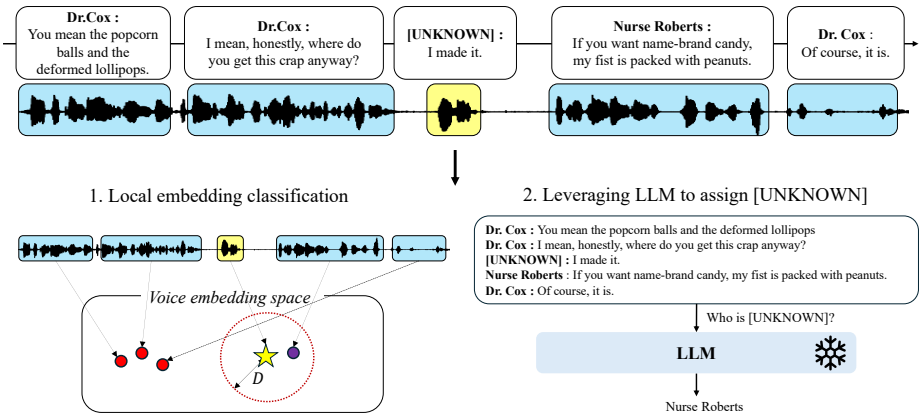

Fig. 2: Assigning speakers to short audio segments. First, we use speaker embeddings from nearby segments where we have high confidence in speaker identification (left). Second, we employ a Large Language Model (LLM) to determine the speaker based on the content of conversation. (right)

It is well known that identifying speakers by their voice alone typically fails in verification tasks when the input audios are short [36, 60]. This is because state-of-the-art speaker embedding extractors [24, 43, 72] are normally trained with segments of at least 2 seconds of audio waveforms. TV shows contain many short segments (see Fig. 5), resulting in false classification when using standard classification methods on the embedding, such as nearest class centroid.

To solve this short segment speaker assignment problem, we use the *temporal context* of the human conversation. There is a high chance that the speaker of the short speech segment we are interested in is involved in the dialogue : which means that the speaker might speak elsewhere and for longer within the scene. The key concept is that speaker identity can be accurately predicted for longer audio segments. These identified segments can then be used to classify speakers in shorter audio segments nearby. We employ this idea into two complementary ways: using local speech embeddings, and using the language (text) of the dialogue. Assuming we know the speakers of long audio segments with a high confidence, we demonstrate how to leverage this information to determine the speakers in shorter audio segments. The method of using temporal context is illustrated in Fig. 2.

Local embedding classification. As is known from diarisation, there are advantages in comparing to *local* embeddings when deciding if two speakers are the same or not. Since the two embeddings are computed under the same environment – and so have the same background sounds, the same reverberation, even the same microphone, many of the 'nuissance' variables are removed, simplifying the classification challenge. Thus, to determine the speaker of a short segment,

its embedding is compared to the segment embeddings of the other (known) speakers in the scene, instead of comparing to their class centroid.

In detail, we extract the speaker embeddings within n_{local} preceding and succeeding sentences around the segment of interest (where $n_{local} = 15$). Then the speaker is assigned by computing the distance between the embeddings of the short segment and segments with known speakers using first nearest neighbor classification. If the distance is below a threshold D then the assignment is accepted, otherwise the short segment is classed as `unknown`, and the assignment is determined (if possible) by using the text content, as described next.

Leverage LLM to assign speakers of `unknown`. As illustrated in Fig. 1, the speaker identity can be inferred solely by using the *content* of the dialogue (*i.e.* without actually hearing the voice). Since large language models (LLMs) have a good predictive 'understanding' of dialogues, they can be queried to predict the speaker of the short segment, given the named speakers of other utterances in the dialogue. We apply this LLM classification in the cases that cannot be classified using voice alone, since it is a weaker cue.

Specifically, we ask the LLM model to predict who the speaker is of the short segment, using zero-shot prompting. We provide the n_{llm} (*e.g.* 15) sentences with the speaker names both before and after this `unknown` sentence. The LLM model is tasked to answer with: either one of the characters that appear within this dialogue with $2n_{llm} + 1$ sentences; or `unknown` if the speaker is from outside the dialogue or if the speaker cannot be inferred only from the provided dialogue. The prompt used also has three examples along with their answers, followed by the actual query and dialogue. The detailed prompt instruction is provided in the supplementary material.

4 Using Local Visual Predictions to Assign Speakers

The task here is to recognise all characters speaking within a video clip. We assume that we have a gallery/library of visual embeddings available for the principal characters. Although speech is sometimes difficult to recognise due to background noise or overlapping voices, the corresponding visual frames often provide a clear view of the speaker. We can use this visual information to help identify speakers. Fig. 3 illustrates the method.

To identify all visible speakers in the scene, we employ a multi-step process. First, we run a pretrained audio-visual synchronisation model [4] that detects lip motions by producing a heat map where areas around moving lips are activated. We then crop spatial regions around the detected peaks with a fixed width and height, and extract visual embeddings from each cropped region using a CLIP-based character recognition model [45]. We compare the distances between these visual embeddings to all actors in the cast list. Finally, we select the cropped regions that identify speakers with high confidence (above a predetermined threshold) and store these predictions for subsequent speaker assignment steps.

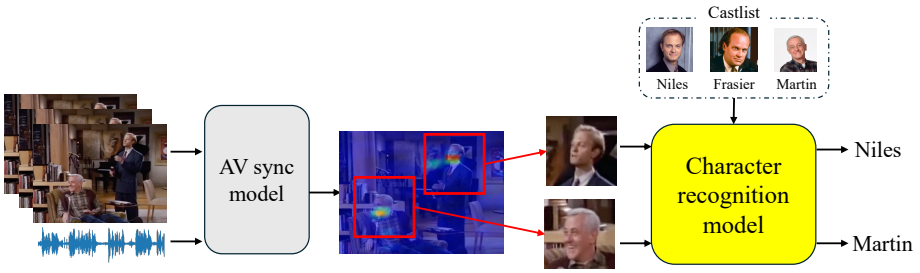

Fig. 3: The visual prediction process for a speech segment. Visible speakers with lip movements synchronised with the speech audio are recognised by using a visual embedding from the castlist. This assigns an identity to the corresponding speaker.

This cropping approach essentially extracts a local visual embedding of the face. It overcomes a limitation of the character recognition model [45], which is confused if multiple people appear in the same frame since it uses a global frame embedding. In summary, we use audio-visual cues to assign identities to speaking segments with visible faces, and audio-only embeddings to assign identities when the face is not visible.

5 Implementation Details

We follow the approach of [44] for generating character-aware subtitles using video and a cast list for each show. Their two-stage method first builds a gallery of audio exemplars – speech segments with high-confidence character name assignments. These exemplars are then used to assign speaker names to all speech segments using centroid classification. If the minimum distance from the nearest exemplar exceeds a threshold, then no specific character name is assigned, allowing for characters without exemplars who cannot be classified. We detail our implementation of this method in the following subsections, highlighting the improvements we have made over the original method of [44]. Fig. 4 shows a schematic overview of our entire pipeline.

5.1 Stage 1. Building audio exemplars

The goal of this stage is to extract audio exemplars from the video for which we know the corresponding speaker.

We first run Voice Activity Detection (VAD) based on Automatic Speech Recognition (ASR) model on the audios to generate the speech transcripts with corresponding timestamps at a sentence level. Visible speakers are then determined by using the synchronisation of lip movements and the speech with the self-supervised trained audio-visual model [5] that produces a heatmap of where the lip-motions are synchronised. We crop the surrounding spatial regions of each peaks in the heatmap and visually recognise characters in the region. Video clips

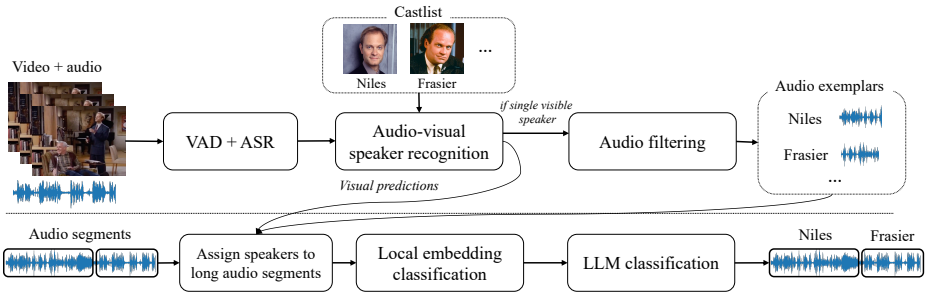

Fig. 4: A schematic overview of our pipeline. We first extract the audio exemplars from videos (top) and use them label all audio segments (bottom).

with a single peak are kept as exemplar candidates, but predictions from the clips with more than one peak are also kept for assigning the speaker later. Then, we conduct additional audio filtering for the exemplar candidates to reduce the label noise. We detail the process below.

Stage 1–1. VAD + ASR. The goal of this stage is to generate speech transcripts with corresponding timestamps. We use publicly available pipeline [10] to produce speech transcripts with timestamps in a sentence level. We assume each sentence is spoken by a single speaker at this point, but we address the case of overlapping speech in subsequent stages. This step produces subtitles without speaker identities. Unlike [44], we do not use any pretrained laughter detector. Instead, we run a speech enhancement network to reduce background noise in the following step.

Stage 1–2. Audio-visual speaker recognition. This stage aims to recognise all speakers in the visual scenes and collect video clips with a single visible speaker using a castlist per video. First, we run a pretrained audio-only speech enhancement model [23] to reduce background noise, thereby reducing false positives in the following stage. Then, we visually recognise the visible speakers' identities by using the method explained in Sec. 4. We need a gallery of images to compare the distance between each visible person and characters in the castlist. We collect up to 10 images per each character and form a visual embedding per character.

After running this model, we categorise the video clips into three types: (i) clips without any peaks, (ii) clips with a single peak, and (iii) clips with multiple peaks. The second type, clips with a single peak, are considered as our exemplar candidates in subsequent stages. However, we proceed to run the character recognition model on both the second and third types. We keep the output predictions to use as candidates for classification in Stage 2. Multiple peaks can occur for two reasons: either multiple speakers are talking simultaneously, or the model produces false positives.

Stage 1–3. Audio filtering This stage further reduces label noise by focusing on single-speaker video clips from the previous stage. We extract speaker embeddings from the audio and analyse each embedding's N ($e.g.$ 5) nearest neighbors. We retain the embedding only if all N neighbors belong to the same speaker; otherwise, we remove the corresponding audio segment from our exemplars.

5.2 Stage 2. Assigning speaker identities of each speech segment

Stage 1 aims to collect audio exemplars for which we know the corresponding speaker identities with high certainty. This stage aims to assign speaker identities to all segments. We first classify the long audio segments (> 2 sec) and segments with extreme high confidence. For each segments, we only compare the distance between the exemplars from the visible speakers, which we obtain in Stage 1–2. If no visible speakers are detected, we compare the distance between the exemplars from castlist. Then, we use the local temporal context, using local embedding classification and LLM which are described in Sec. 3. The long segments as well as audio exemplars from the previous stage are used for local embedding classification.

After assigning speakers to each audio segments in a sentence level, we run the public overlapping speech detection model to detect the overlapping speech. If the segment is detected with overlapping speech, we assign the speaker with two nearest speakers along the time axis.

5.3 Implementation details

We use WhisperX [10] for VAD+ASR model. We further use Silero VAD [71] for filtering out the false detections from the WhisperX. ECAPA-TDNN [24] is used for speaker embedding extractor, pretrained with VoxCeleb [56]. We use LWTNet [5] for audio-visual synchronisation model and crop the activate region with $W = H = 350px$ to recognise the characters. We use $n_{local} = 15$ preceding and suceeding audio segments. We use public official Llama3-70B 4-bit quantised model, finetuned with instruction sets, to assign speakers with $n_{LLM} = 15$ preceding and succeeding sentences. We use public overlapping detection model from `pyannote 2.1` [15]. Rest of the parameters are identical to those in [44]. After detecting overlapping speech, we divide the audio segments wherever there is silence longer than 1 second, using word-level timestamps from WhisperX. The same speaker is assigned to these divided audio segments. All hyperparameters are determined by grid search on validation sets.

6 Dataset and evaluation metrics

This section explains the dataset we have used to validate our method and evaluation metrics.

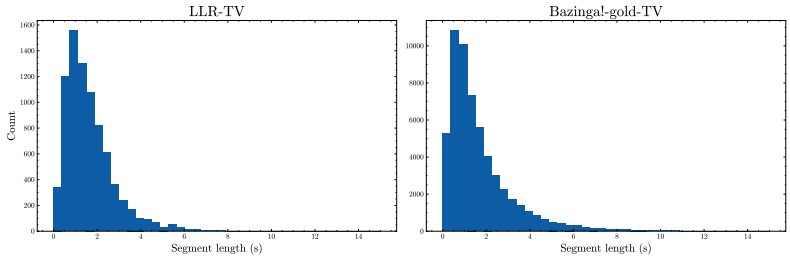

Fig. 5: Distribution of segment lengths on LLR-TV and *Bazinga!*-gold-TV.

6.1 Dataset

LLR-TV: [44] has released a TV shows dataset including six episodes each from Frasier, Seinfeld, and Scrubs, along with transcripts, speaker names, and timestamps. The official dataset website [1] has released version 1.1, which includes fixed annotations they have made. We use the current version to verify our method, but we also report the performance from the original paper in Sec. 7. For each series, we use the sixth episode as our validation set to determine the hyperparameters. The rest of the dataset is used as our test set.

***Bazinga!*-gold-TV:** *Bazinga!* [48] dataset provides a rich set of annotations from 16 different TV shows and movies, such as speech transcripts with timestamps, speaker, addressee and entity linking information. The dataset itself is divided into gold and silver based on the level of annotations. We use all TV shows in the gold set to verify our method including *Battlestar Galactica (B.G.)*, *Breaking Bad (B.B.)*, *Buffy the Vampire Slayer (Buffy)*, *Friends*, *Game of Thrones (GoT)*, *Lost*, *The Big Bang Theory (TBBT)*, *The Office (Office)* and *The Walking Dead (W.D.)*. We exclude *StarWars* since our paper focuses on TV shows.

Since our method is audio-visual while the dataset only provides audio, we need to adjust the timestamps in the annotations to match our videos. We use the audio-audio alignment method introduced by [34] to obtain precise temporal alignment by comparing the audio provided in the dataset with the audio from our video source. Similar to LLR-TV, we use the last episode of each series as our validation set, with the remaining episodes serving as our test set.

Fig. 5 shows the distribution of segment lengths in both datasets. 71.0% of segments in LLR-TV and 71.5% in *Bazinga!*-gold-TV are shorter than 2 seconds. This indicates that recognising speakers in shorter segments is crucial for analysing conversations in TV shows. Note that while LLR-TV is manually corrected, *Bazinga!*-gold-TV provides timestamps obatined through force-alignment. Thus the annotation is relatively noisy (*e.g.* they do not provide annotations for *Previously...* part.).

[1] `https://www.robots.ox.ac.uk/~vgg/research/look-listen-recognise/`

6.2 Evaluation metrics

Diarisation metrics. DER [58] is a standard evaluation metric for speaker diarisation. However, recent studies [20, 51] have highlighted a significant limitation of DER: its time-duration-based computation fails to accurately capture the recognition performance for short-term segments. To address this limitation, Conversational-DER (**CDER**) is introduced, which calculates speaker diarization accuracy at the utterance level and also accounts for short segments. For more details on CDER, please refer to [20].

In this paper, we employ DER with a forgiveness collar of 0.25 seconds, taking into consideration instances of overlapping speech.

Character recognition metrics. Character recognition accuracy (**Acc.**) is calculated for segments that overlap with ground truth segments. A segment is considered correctly classified if the character's name is accurately identified and matches the corresponding overlapping ground truth segment. Precision and recall for character identification are also reported for each show. Both metrics are calculated for all characters (**Prec.** and **Rec.**) and separately for main characters (**Prec.(M)** and **Rec.(M)**) in each series. A list of main characters for each show is provided in the supplementary material.

7 Results

This section presents our overall results on the test sets, comparing them to other baselines. We also provide a detailed analysis of how our method accurately collect audio exemplar. We conclude by showcasing qualitative examples of our method and comparing them to the baseline. The effect of using local visual predictions and speech enhancement is shown in the supplementary.

7.1 Overall performance

Diarisation performance. We report the diarisation performance of our method on the LLR-TV test set in Tab. 1. Our method is compared against three competitive baselines, including two audio-only models and one audio-visual diarisation method, LLR. In terms of DER, our method demonstrates superior performance on Frasier and Seinfeld compared to all other models, and achieves comparable results on Scrubs. More notably, when considering the CDER, our method significantly outperforms other baselines with margins of 10.3%, 7.1%, and 3.3% on Frasier, Seinfeld, and Scrubs, respectively. This indicates that our method recognises characters more accurately, even in short segments, compared to other methods.

Tab. 2 compares our pipeline against other baselines on the *Bazinga!*-gold-TV, where our method shows better performance in most TV series. The metrics should be taken with a 'grain of salt', a point that is also made in the original paper [48].

Table 1: Diarisation performance on LLR-TV test set. Lower is better. **LLR**$*$ is from the original paper before the GT was corrected and **LLR**† is our reproduced result with annotation corrections from the website. **DER** : Diarisation Error Rate (%), **CDER** : Communication DER (%), **A** : Audio, **V** : Video.

		Frasier		Seinfeld		Scrubs	
Model	**Modality**	**DER**	**CDER**	**DER**	**CDER**	**DER**	**CDER**
SimpleDiar [66]	**A**	24.2	58.5	24.5	56.2	**24.4**	52.6
`pyannote` [14]	**A**	24.7	84.1	35.4	88.7	31.1	75.8
LLR$*$ [44]	**A + V**	24.3	-	29.7	-	36.4	-
LLR† [44]	**A + V**	26.4	39.1	28.0	40.7	26.7	40.3
Ours	**A + V**	**20.3**	**28.8**	**23.3**	**33.6**	25.7	**37.0**

Table 2: Diarisation performance on *Bazinga!*-gold-TV. **DER** : Diarisation Error Rate (%), **CDER** : Communication DER (%), **Mod** : Modality, **A** : audio, **V** : Video.

		B.G.		B.B.		Buffy		Friends		GoT		Lost		TBBT		Office		W.D.	
Model	Mod	DER	CDER	DER	CDER	DER	CDER	DER	CDER	DER	CDER	DER	CDER	DER	CDER	DER	CDER	DER	CDER
SimpleDiar [66]	**A**	61.3	101.0	68.8	154.5	31.5	62.6	**46.8**	85.6	38.3	85.8	90.9	117.2	**20.6**	38.0	**33.3**	**70.3**	93.4	123.9
`pyannote` [14]	**A**	**58.7**	104.9	70.4	136.7	**31.3**	**59.6**	60.5	128.8	**37.2**	85.1	**88.1**	111.9	30.3	70.0	35.4	112.5	100.7	138.2
LLR† [44]	**A+V**	79.5	101.6	92.9	135.4	55.4	77.5	55.9	72.6	63.7	120.0	111.7	115.4	29.5	39.5	44.2	93.9	108.9	130.6
Ours	**A+V**	62.7	**87.2**	**67.0**	**99.4**	46.2	62.0	47.1	**64.4**	44.2	**82.7**	89.0	**86.6**	27.3	**36.0**	40.2	71.8	**92.5**	**97.1**

Character recognition performance. Tab. 3 shows the character recognition performance on LLR-TV. The reported performance from both LLR and our method is based on the highest accuracy achieved by varying the hyperparameter D (see Sec. 3) on the validation set. Compared to the reproduced LLR, our method demonstrates higher character recognition accuracy. Interestingly, although the LLM is instructed not to predict the speaker when it cannot be inferred from the input dialogue, it mostly selects a speaker from within the dialogue, resulting in higher recall for both all segments and segments from main characters. LLR does not predict speakers for 6.8% of the test set, while our method classifies only 0.77% as `unknown`.

Precision-POCS curve. We demonstrate the Precision – Proportion of Classified Segments (POCS) trade-off in Fig. 6, showing results for all segments (left), long segments (middle), and short segments (right) by varying the threshold D. We include the curve from the LLR method for comparison. The graphs show that the precision of character recognition decreases as we classify more audio segments. Compared to LLR, our method shows similar performance on long segments but demonstrates superior ability in classifying short segments. This verifies our method's effectiveness in identifying speakers of short utterances.

7.2 Exemplar recognition accuracy

Tab. 4 shows the accuracy of character recognition for the exemplars. We demonstrate that the accuracy of the audio exemplar building stage is nearly perfect.

Table 3: Character recognition performance on LLR-TV test set. **Prec.** and **Rec.** indicate the precision and recall of overall audio segments respectively, while **Prec.(M)** and **Rec.(M)** are those of main characters in TV shows. **Acc.** is the character recognition accuracy for those which overlap with one of the groundtruth timestamps.

	LLR [44]					**Ours**				
	Acc.	**Prec.**	**Rec.**	**Prec.(M)**	**Rec.(M)**	**Acc.**	**Prec.**	**Rec.**	**Prec.(M)**	**Rec.(M)**
Frasier	87.0	**91.6**	87.0	**92.5**	89.4	**88.9**	89.8	**88.9**	90.3	**92.6**
Seinfeld	84.5	**89.0**	84.6	**92.5**	89.4	**85.8**	87.1	**86.0**	89.5	**90.7**
Scrubs	84.3	**89.2**	84.9	**91.0**	88.1	**84.4**	84.8	**85.1**	85.1	**90.7**

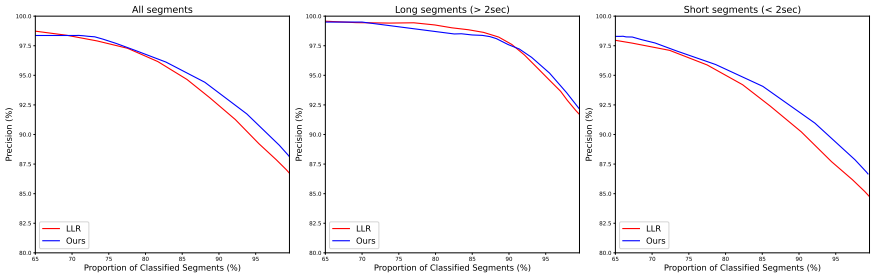

Fig. 6: Precision-POCS curves for audio segments in LLR-TV test set.

Out of 1,734 exemplars, most characters show 100% accuracy. The pipeline mispredicts only **6** speakers (0.34%), 5 from Frasier and one from Seinfeld. This high accuracy offers two advantages: (i) more audio segments are correctly classified, and (ii) more precise embeddings are obtained for local embedding classification of short segments. In terms of the number of exemplars, we extract more audio samples than [44] from the same Seinfeld test set (609 vs. 407).

Table 4: Exemplar recognition performance on LLR-TV. # **exem** denotes the number exemplars extracted from our method and # **correct** denotes the number of correctly classified exemplars. **Others** are a group of characters who are usually guest stars for one or a few episodes and the number of them is given in parentheses.

Frasier				**Seinfeld**				**Scrubs**			
Char	**# exem**	**# correct**	**Acc (%)**	**Char**	**# exem**	**# correct**	**Acc (%)**	**Char**	**# exem**	**# correct**	**Acc (%)**
Frasier	347	342	98.6	**Jerry**	353	352	99.7	**J.D.**	152	152	100
Niles	58	58	100	**George**	41	41	100	**Dr.Cox**	102	102	100
Roz	28	28	100	**Elaine**	66	66	100	**Turk**	22	22	100
Daphne	30	30	100	**Kramer**	27	27	100	**Dr.Kelso**	65	65	100
Martin	31	31	100					**Elliot**	88	88	100
								Carla	73	73	100
Others (7)	47	47	100	**Others (12)**	122	122	100	**Others (13)**	82	82	100

Fig. 7: Qualitative examples from two TV series, *Scrubs* and *Friends*.

7.3 Qualitative examples

We present qualitative results from two series, *Scrubs* and *Friends*, in Fig. 7. The figure shows speech recognition output and corresponding timestamps produced by our method, along with character recognition results from both LLR and this approach. In both series, LLR fails to predict speakers for short utterances such as *"You know."*, *"Me."* or *"Why not?"*. In contrast, our method utilises the temporal context of the conversation to correctly classify the speaker for these brief segments. It is important to note that the yellow utterances in the figure are initially classified as `unknown`. However, after employing LLM, these are correctly assigned to the appropriate speakers.

8 Conclusions

This paper presents an advanced framework for character-aware audio-visual subtitling in TV shows, addressing limitations in existing methods. Key contributions include a novel method for identifying speakers in short segments using temporal context, the use of local visual embeddings around lip-moving areas, and validation on a large dataset covering 12 TV series. Results demonstrate significant improvements in both diarisation performance and character recognition accuracy, particularly for short speech segments.

Acknowledgments. This research is supported by EPSRC Programme Grant VisualAI EP/T028572/1 and a Royal Society Research Professorship RP\R1\191132. We thank Robin and Bruno for helpful discussions.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023)
2. Afouras, T., Chung, J.S., Senior, A., Vinyals, O., Zisserman, A.: Deep audio-visual speech recognition. IEEE PAMI (2019)
3. Afouras, T., Chung, J.S., Zisserman, A.: Lrs3-ted: a large-scale dataset for visual speech recognition. arXiv preprint arXiv:1809.00496 (2018)
4. Afouras, T., Chung, J.S., Zisserman, A.: Now you're speaking my language: Visual language identification. In: INTERSPEECH (2020)
5. Afouras, T., Owens, A., Chung, J.S., Zisserman, A.: Self-supervised learning of audio-visual objects from video. In: Proc. ECCV (2020)
6. AI@Meta: Llama 3 model card (2024), `https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md`
7. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., et al.: Flamingo: a visual language model for few-shot learning. Advances in neural information processing systems **35**, 23716–23736 (2022)
8. Baevski, A., Zhou, Y., Mohamed, A., Auli, M.: wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems **33**, 12449–12460 (2020)
9. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv preprint arXiv:2308.12966 (2023)
10. Bain, M., Huh, J., Han, T., Zisserman, A.: Whisperx: Time-accurate speech transcription of long-form audio. In: INTERSPEECH (2023)
11. Berg, T., Berg, A., Edwards, J., Mair, M., White, R., Teh, Y., Learned-Miller, E., Forsyth, D.: Names and Faces in the News. In: Proc. CVPR (2004)
12. Bojanowski, P., Bach, F., , Laptev, I., Ponce, J., Schmid, C., Sivic, J.: Finding actors and actions in movies. In: Proc. ICCV (2013)
13. Bost, X., Linarès, G., Gueye, S.: Audiovisual speaker diarization of tv series. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4799–4803. IEEE (2015)
14. Bredin, H.: pyannote. audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In: Proc. Interspeech. pp. 1983–1987. ISCA (2023)
15. Bredin, H., Laurent, A.: End-to-end speaker segmentation for overlap-aware resegmentation. In: Proc. Interspeech 2021. Brno, Czech Republic (August 2021)
16. Brown, A., Coto, E., Zisserman, A.: Automated video labelling: Identifying faces by corroborative evidence. In: International Conference on Multimedia Information Processing and Retrieval (2021)
17. Brown, A., Kalogeiton, V., Zisserman, A.: Face, body, voice: Video person-clustering with multiple modalities. In: ICCV 2021 Workshop on AI for Creative Video Editing and Understanding (2021)
18. Chan, W., Jaitly, N., Le, Q.V., Vinyals, O.: Listen, attend and spell. arXiv preprint arXiv:1508.01211 (2015)
19. Chen, J., Zhu, D., Haydarov, K., Li, X., Elhoseiny, M.: Video chatcaptioner: Towards the enriched spatiotemporal descriptions. arXiv preprint arXiv:2304.04227 (2023)

20. Cheng, G., Chen, Y., Yang, R., Li, Q., Yang, Z., Ye, L., Zhang, P., Zhang, Q., Xie, L., Qian, Y., et al.: The conversational short-phrase speaker diarization (cssd) task: Dataset, evaluation metric and baselines. In: 2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP). pp. 488–492. IEEE (2022)
21. Chung, J.S., Huh, J., Mun, S., Lee, M., Heo, H.S., Choe, S., Ham, C., Jung, S., Lee, B.J., Han, I.: In defence of metric learning for speaker recognition. In: Interspeech (2020)
22. Chung, J.S., Huh, J., Nagrani, A., Afouras, T., Zisserman, A.: Spot the conversation: speaker diarisation in the wild. In: INTERSPEECH (2020)
23. Defossez, A., Synnaeve, G., Adi, Y.: Real time speech enhancement in the waveform domain. In: Interspeech (2020)
24. Desplanques, B., Thienpondt, J., Demuynck, K.: ECAPA-TDNN: Emphasized Channel Attention, propagation and aggregation in TDNN based speaker verification. In: Interspeech 2020. pp. 3830–3834 (2020)
25. Diez, M., Burget, L., Landini, F., Černockỳ, J.: Analysis of speaker diarization based on bayesian hmm with eigenvoice priors. IEEE/ACM Transactions on Audio, Speech, and Language Processing **28**, 355–368 (2019)
26. Everingham, M., Sivic, J., Zisserman, A.: "Hello! My name is... Buffy" – automatic naming of characters in TV video. In: Proc. BMVC (2006)
27. Everingham, M., Sivic, J., Zisserman, A.: Taking the bite out of automatic naming of characters in TV video. Image and Vision Computing **27**(5) (2009)
28. Fujita, Y., Kanda, N., Horiguchi, S., Nagamatsu, K., Watanabe, S.: End-to-End Neural Speaker Diarization with Permutation-free Objectives. In: Interspeech. pp. 4300–4304 (2019)
29. Gabeur, V., Seo, P.H., Nagrani, A., Sun, C., Alahari, K., Schmid, C.: Avatar: Unconstrained audiovisual speech recognition. arXiv preprint arXiv:2206.07684 (2022)
30. Gong, Y., Liu, A.H., Luo, H., Karlinsky, L., Glass, J.: Joint audio and speech understanding. In: 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) (2023)
31. Gong, Y., Luo, H., Liu, A.H., Karlinsky, L., Glass, J.: Listen, think, and understand. In: Proc. ICLR (2023)
32. Gulati, A., Qin, J., Chiu, C.C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., et al.: Conformer: Convolution-augmented transformer for speech recognition. arXiv preprint arXiv:2005.08100 (2020)
33. Han, T., Bain, M., Nagrani, A., Varol, G., Xie, W., Zisserman, A.: Autoad: Movie description in context. In: Proc. CVPR (2023)
34. Han, T., Bain, M., Nagrani, A., Varol, G., Xie, W., Zisserman, A.: AutoAD III: The prequel – back to the pixels. In: Proc. CVPR (2024)
35. Haurilet, M.L., Tapaswi, M., Al-Halah, Z., Stiefelhagen, R.: Naming tv characters by watching and analyzing dialogs. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1–9. IEEE (2016)
36. He, Y., Kang, Z., Wang, J., Peng, J., Xiao, J.: Voiceextender: Short-utterance text-independent speaker verification with guided diffusion model. In: 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU). pp. 1–8. IEEE (2023)
37. Hsu, W.N., Bolte, B., Tsai, Y.H.H., Lakhotia, K., Salakhutdinov, R., Mohamed, A.: Hubert: Self-supervised speech representation learning by masked prediction of hidden units. IEEE/ACM Transactions on Audio, Speech, and Language Processing **29**, 3451–3460 (2021)

38. Hu, Y., Ren, J.S., Dai, J., Yuan, C., Xu, L., Wang, W.: Deep multimodal speaker naming. In: Proceedings of the 23rd ACM international conference on Multimedia. pp. 1107–1110 (2015)
39. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.d.l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., et al.: Mistral 7b. arXiv preprint arXiv:2310.06825 (2023)
40. Kalogeiton, V., Zisserman, A.: Constrained video face clustering using 1nn relations. In: Proc. BMVC (2020)
41. Kaphungkui, N., Kandali, A.B.: Text dependent speaker recognition with back propagation neural network. International Journal of Engineering and Advanced Technology (IJEAT) **8**(5), 1431–1434 (2019)
42. Kinoshita, K., Delcroix, M., Tawara, N.: Integrating end-to-end neural and clustering-based diarization: Getting the best of both worlds. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 7198–7202. IEEE (2021)
43. Koluguri, N.R., Park, T., Ginsburg, B.: Titanet: Neural model for speaker representation with 1d depth-wise separable convolutions and global context. In: ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 8102–8106. IEEE (2022)
44. Korbar, B., Huh, J., Zisserman, A.: Look, listen and recognise: character-aware audio-visual subtitling. In: International Conference on Acoustics, Speech, and Signal Processing (2024)
45. Korbar, B., Zisserman, A.: Personalised clip or: how to find your vacation videos. In: Proc. BMVC (2022)
46. Kraaij, W., Hain, T., Lincoln, M., Post, W.: The ami meeting corpus. In: Proc. International Conference on Methods and Techniques in Behavioral Research (2005)
47. Kwak, D., Jung, J., Nam, K., Jang, Y., Jung, J.w., Watanebe, S., Chung, J.S.: Voxmm: Rich transcription of conversations in the wild. In: International Conference on Acoustics, Speech, and Signal Processing (2024)
48. Lerner, P., Bergoënd, J., Guinaudeau, C., Bredin, H., Maurice, B., Lefevre, S., Bouteiller, M., Berhe, A., Galmant, L., Yin, R., et al.: Bazinga! a dataset for multi-party dialogues structuring. In: 13th Conference on Language Resources and Evaluation (LREC 2022). pp. 3434–3441 (2022)
49. Li, K., Wrench Jr, E.: Text-independent speaker recognition with short utterances. The Journal of the Acoustical Society of America **72**(S1), S29–S30 (1982)
50. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023)
51. Liu, T., Yu, K.: Ber: Balanced error rate for speaker diarization. arXiv preprint arXiv:2211.04304 (2022)
52. Lu, H., Liu, W., Zhang, B., Wang, B., Dong, K., Liu, B., Sun, J., Ren, T., Li, Z., Sun, Y., Deng, C., Xu, H., Xie, Z., Ruan, C.: Deepseek-vl: Towards real-world vision-language understanding (2024)
53. Ma, P., Haliassos, A., Fernandez-Lopez, A., Chen, H., Petridis, S., Pantic, M.: Auto-avsr: Audio-visual speech recognition with automatic labels. In: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 1–5. IEEE (2023)
54. Maaz, M., Rasheed, H., Khan, S., Khan, F.S.: Video-chatgpt: Towards detailed video understanding via large vision and language models. In: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024) (2024)
55. Malik, M., Malik, M.K., Mehmood, K., Makhdoom, I.: Automatic speech recognition: a survey. Multimedia Tools and Applications **80**, 9411–9457 (2021)

56. Nagrani, A., Chung, J.S., Xie, W., Zisserman, A.: Voxceleb: Large-scale speaker verification in the wild. Computer Speech and Language (2019)
57. Nagrani, A., Zisserman, A.: From benedict cumberbatch to sherlock holmes: Character identification in tv series without a script. In: Proc. BMVC (2017)
58. The 2009 (RT-09) Rich Transcription Meeting Recognition Evaluation Plan (2009 (accessed 1 July 2024)), https://web.archive.org/web/20100606092041if_/http://www.itl.nist.gov/iad/mig/tests/rt/2009/docs/rt09-meeting-eval-plan-v2.pdf, See Section 6
59. Park, T.J., Kanda, N., Dimitriadis, D., Han, K.J., Watanabe, S., Narayanan, S.: A review of speaker diarization: Recent advances with deep learning. Computer Speech & Language **72**, 101317 (2022)
60. Poddar, A., Sahidullah, M., Saha, G.: Speaker verification with short utterances: a review of challenges, trends and opportunities. IET Biometrics **7**(2), 91–101 (2018)
61. Poignant, J., Bredin, H., Barras, C.: Multimodal person discovery in broadcast tv: lessons learned from mediaeval 2015. Multimedia Tools and Applications **76**, 22547–22567 (2017)
62. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: International Conference on Machine Learning. pp. 28492–28518. PMLR (2023)
63. Ramanathan, V., Joulin, A., Liang, P., Fei-Fei, L.: Linking people in videos with "their" names using coreference resolution. In: Proc. ECCV. pp. 95–110. Springer (2014)
64. Sharma, R., Narayanan, S.: Using active speaker faces for diarization in tv shows. arXiv preprint arXiv:2203.15961 (2022)
65. Shi, B., Hsu, W.N., Lakhotia, K., Mohamed, A.: Learning audio-visual speech representation by masked multimodal cluster prediction. arXiv preprint arXiv:2201.02184 (2022)
66. Simple diarization repository. https://github.com/JaesungHuh/SimpleDiarization (2024)
67. Song, E., Chai, W., Wang, G., Zhang, Y., Zhou, H., Wu, F., Guo, X., Ye, T., Lu, Y., Hwang, J.N., et al.: Moviechat: From dense token to sparse memory for long video understanding. In: Proc. CVPR (2024)
68. Suchitha, T., Bindu, A.: Feature extraction using mfcc and classification using gmm. International Journal for Scientific Research & Development (IJSRD) **3**(5), 1278–1283 (2015)
69. Szarkowska, A.: Subtitling for the deaf and the hard of hearing. The Palgrave Handbook of Audiovisual Translation and Media Accessibility pp. 249–268 (2020)
70. Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023)
71. Team, S.: Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier. https://github.com/snakers4/silero-vad (2021)
72. Torgashov, N., Makarov, R., Yakovlev, I., Malov, P., Balykin, A., Okhotnikov, A.: The id r&d voxceleb speaker recognition challenge 2023 system description. arXiv preprint arXiv:2308.08294 (2023)
73. Wang, J., Chen, D., Luo, C., Dai, X., Yuan, L., Wu, Z., Jiang, Y.G.: Chatvideo: A tracklet-centric multimodal and versatile video understanding system. arXiv preprint arXiv:2304.14407 (2023)

74. Wang, Q., Downey, C., Wan, L., Mansfield, P.A., Moreno, I.L.: Speaker diarization with lstm. In: 2018 IEEE International conference on acoustics, speech and signal processing (ICASSP). pp. 5239–5243. IEEE (2018)
75. Xu, E.Z., Song, Z., Tsutsui, S., Feng, C., Ye, M., Shou, M.Z.: Ava-avd: Audio-visual speaker diarization in the wild. In: Proceedings of the 30th ACM International Conference on Multimedia. pp. 3838–3847 (2022)
76. Zhang, A., Wang, Q., Zhu, Z., Paisley, J., Wang, C.: Fully supervised speaker diarization. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6301–6305. IEEE (2019)
77. Zhang, H., Li, X., Bing, L.: Video-llama: An instruction-tuned audio-visual language model for video understanding. arXiv preprint arXiv:2306.02858 (2023), https://arxiv.org/abs/2306.02858