Supplementary material for "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

1 Character recognition accuracy on *Bazinga!*-gold-TV

Tab. 1 shows the character recognition performance on *Bazinga!*-gold-TV, comparing with that from LLR [1]. The reported performance of LLR corresponds to its highest character recognition accuracy (**Acc.**). This peak accuracy is achieved by adjusting the threshold during the nearest centroid classification process. We showcase that our method achieves higher accuracy than LLR in all TV shows. It also achieves higher recall for both 'all characters' and 'main characters' compared to LLR. LLR exhibits higher precision for 'all characters' and 'main characters' in a few TV shows. This difference in performance is due to how each method handles short segments. LLR does not assign speakers to some of short segments, instead marking the character as `unknown`. In contrast, our LLM tends to predict speakers for these `unknown` segments after local embedding classification. It often assigns one of the characters appearing in the dialogue, even when we explicitly instruct the LLM that it doesn't have to make an assignment if it's unsure. Thus, the precision decreases but the recall increases.

2 Effects of utilising spatial regions and speech enhancement on audio exemplar yield and performance

Tab. 2 demonstrates how cropping lip areas and speech enhancement improve exemplar yield. Unlike LLR, which uses whole frames without speech enhancement, our method identifies characters by focusing on the spatial region of speakers and reducing background noise to decrease false positive peaks, increasing exemplar yield by 8.5% and 0.9% respectively. This showcases the effectiveness of our cropping-based approach in terms of exemplar yield.

3 List of main characters per series

Tab. 3 lists the main characters per series, both in LLR-TV and *Bazinga!*-gold-TV. Note that we report the names of main characters in *Bazinga!*-gold-TV as they are in the dataset annotation.

Table 1: Character recognition performance on *Bazinga!*-gold-TV test set. **Acc.** is the character recognition accuracy for those which overlap with one of the groundtruth timestamps. **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.

	LLR [1]					**Ours**				
	Acc.	**Prec.**	**Rec.**	**Prec.(M)**	**Rec.(M)**	**Acc.**	**Prec.**	**Rec.**	**Prec.(M)**	**Rec.(M)**
B.G.	62.7	67.7	64.1	**72.7**	71.1	**68.7**	**69.1**	**70.1**	69.1	**75.9**
B.B.	60.5	69.4	61.4	**83.6**	64.9	**67.5**	**72.6**	**68.6**	76.3	**69.4**
Buffy	57.6	62.4	58.1	71.7	57.5	**61.6**	**62.6**	**62.1**	**75.0**	**65.0**
Friends	70.9	**82.1**	72.9	**85.8**	79.7	**74.2**	75.6	**76.4**	76.0	**83.2**
GoT	52.6	55.4	55.7	**57.3**	59.7	**61.6**	**62.9**	**65.3**	54.1	**63.7**
Lost	51.4	61.2	53.4	62.6	56.6	**60.4**	**63.9**	**62.6**	**65.4**	**67.3**
TBBT	80.2	**88.8**	80.4	**90.9**	85.4	**81.5**	82.7	**81.8**	83.2	**86.2**
Office	68.7	**77.5**	68.9	**84.8**	74.1	**70.8**	74.4	**71.2**	75.4	**75.0**
W.D.	49.7	**60.3**	51.9	**57.0**	52.9	**54.7**	57.6	**57.1**	53.5	**64.0**

Table 2: Effect of using local visual predictions (vis) and speech enhancement (SE) on exemplar yield on the LLR-TV.

	LLR		LLR + vis		LLR + vis + SE (Ours)	
	# of exemplars	% of total	# of exemplars	% of total	# of exemplars	% of total
1-1. VAD + ASR	5554	100	5554	100	5554	100
1-2. Audio-visual speaker recognition	2192	39.5	3332	60.0	3409	**61.4**
1-3. Audio filtering	1213	21.8	1681	30.3	1734	**31.2**

4 LLM prompt

Tab. 4 shows the LLM prompt we've used to determine the `unknown` character in the dialogue. We adopt the strategy introduced in [2]. Given the query dialogue with the `unknown` character we want to classify, we first ask the LLM to summarize the dialogue. Then, we ask the model who would be the `unknown` in the dialogue based on the generated summary. A list of characters with their corresponding indices is provided in the prompt. We compute the softmaxed logits of the tokens corresponding to each index and choose the largest one to assign the speaker. We use this same prompt for experiments in LLR-TV and *Bazinga!*-gold-TV.

There can be multiple `unknown`s in the query dialogue. We mark the sentence for which we want to identify the speaker by placing asterisks (**) before and after it. This explicitly instructs the LLM to predict the speaker only for that specific sentence.

References

1. Korbar, B., Huh, J., Zisserman, A.: Look, listen and recognise: character-aware audio-visual subtitling. In: International Conference on Acoustics, Speech, and Sig-

Table 3: List of main characters per show.

Show	Main characters
LLR-TV	
Frasier	Frasier, Martin, Niles, Roz, Daphne
Seinfeld	Jerry, Elaine, George, Kramer
Scrubs	J.D., Dr.Cox, Dr.Kelso, Carla, Turk, Elliot
***Bazinga!*-gold-TV**	
Battlestar Galactica (B.G.)	Admiral William Adama, President Laura Roslin, Captain Kara Thrace, Cpt. Lee Adama, Number six, Dr. Gaius Baltar, Lt. Sharon Valerii
Breaking Bad (B.B.)	Walter White, Skyler White, Jesse Pinkman, Hank Schrader, Marie Schrader
Buffy the Vampire Slayer (Buffy)	Buffy Summers, Willow Rosenberg, Xander Harris, Angel, Rubert Giles
Friends	Rachel Green, Monica Geller, Phoebe Buffay, Joey Tribbiani, Chandler Bing, Dr. Ross Geller
Game of Thrones (GoT)	Danerys Targaryen, Jon Snow, Jorah Mormont, Tyrion Lannister, Catelyn Stark, Sansa Stark, Arya Stark, Cersei Lannister, Eddard Stark, Robert Baratheon
Lost	Dr. Jack Sheperd, Kate Austen, Sayid Jarrah, Hugo Reyes, Sunhwa Kwon, Charlie Pace, Clarie Littleton, Michael Dawson, John Locke, Shannon Rutherford, James Ford
The Big Bang Theory (TBBT)	Sheldon Cooper, Penny, Howard Wolowitz, Raj Koothrappali, Leonard Hofstadter
The Office (Office)	Jim Halpert, Michael Scott, Ryan Howard, Pam Beesly, Dwight Schrute, Stanley Hudson, Phyllis Vance, Angela Martin
Lost	Rick Grimes, Lori Grimes, Carl Grimes, Carol Peletier, Shane Walsh, Andrea Harrison, Dale Horvath, Glenn Rhee

nal Processing (2024)
2. Park, R.Y., Windsor, R., Jamaludin, A., Zisserman, A.: Automated spinal mri labelling from reports using a large language model. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 101–111. Springer (2024)

Table 4: LLM prompt to determine `unknown`.

Prompt for determining [UNKNOWN]

system : You are a AI assistant to analyze the transcript of TV shows. Your job is to figure out who are [UNKNOWN]s in a dialogue in TV shows. Tell the truth and answer as precisely as possible.
(provide the dialogue here)
user : Write a summary for the above conversation.
assistant : *(model generates the summary)*
user : Based on the summary, your job is to identify the name of the speaker of '[UNKNOWN]' when the line starts and ends with '**'. You must use the context and the flow of the dialogue, using the speakers' names and what they speak. The list of speakers with their corresponding tokens are provided below. Choose [UNKNOWN] if his or her name is not in the dialogue, or when you are not sure.
(provide the list of speakers here)
Only output one number after ANSWER:

Example of the dialogue

Dr.Cox : You've got an opposable thumb.
Dr.Cox : You can use it.
Dr.Cox : God, I hate Halloween.
Carla : Somebody needs to adjust their attitude if they want some candy.
Dr.Cox : You mean, the popcorn balls and the deformed lollipops.
Dr.Cox : I mean, honestly, where do you get this crap anyway?
[UNKNOWN] : I made it.
NurseRoberts : If you want name brand candy, my fish is packed with peanuts.
Dr.Cox : Of course it is.
Carla : Oh, what's the matter?
Carla : Did Raggedy Ann scare you?
Dr.Cox : What are you, a rat?

Example of the list of speakers

1: Dr.Cox, 2: Carla, 3: NurseRoberts, 4: [UNKNOWN]