CI-AVSR: A Cantonese Audio-Visual Speech Dataset for In-car Command Recognition

Wenliang Dai, Samuel Cahyawijaya, Tiezheng Yu, Elham J. Barezi,
Peng Xu, Cheuk Tung Shadow Yiu, Rita Frieske, Holy Lovenia,
Genta Indra Winata, Qifeng Chen, Xiaojuan Ma, Bertram E. Shi, Pascale Fung

The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong

Introduction

We introduce a new dataset, **C**antonese **I**n-car **A**udio-**V**isual **S**peech **R**ecognition (CI-AVSR), for in-car command recognition in the Cantonese language with both video and audio data. **CI-AVSR** consists of **4,984 samples (8.3 hours)** of **200 in-car commands** recorded by **30 native Cantonese speakers**. In addition, we augment our dataset using common in-car background noises, including **alarm**, **horn**, **background music**, **ignition**, **hail**, **rain**, **windscreen wiper**, **road ambience**, and **door opens and closes**, to simulate real environments, producing a dataset 10 times larger than the originally collected one.

Baseline Models

Model	Audio only (CER)		Audio + Video (CER)	
	clean	noisy	clean	noisy
Conformer	31.29%	167.63%	26.97%	132.68%
Wav2Vec2	4.06%	12.75%	3.48%	7.19%

Table 1:Evaluation CER of two baseline models on both the clean and augmented (noisy) test sets.

Type of Noise	CER	
	Audio	Audio + Video
Clean	4.06%	3.48%
0 (background music)	11.53%	5.42%
1 (rain)	12.16%	6.19%
2 (hail)	17.03%	10.03%
3 (ignition)	21.73%	13.65%
4 (windscreen wiper)	13.57%	8.50%
5 (horn)	16.26%	9.06%
6 (people talking)	13.53%	7.78%
7 (road ambience)	13.25%	7.36%
8 (alarm)	17.13%	8.58%
9 (car door)	7.19%	3.38%
Avg (0 to 9)	14.34%	7.99%

Table 2:The character error rate of the Wav2Vec 2.0 model (trained on the clean training set only) on the augmented noisy data. We report its performance on each type of noise and the average of them.

Conclusion

We introduce **Cantonese In-car Audio-Visual Speech Recognition (CI-AVSR)**, for audio-visual speech recognition of in-car commands. It consists of **200 unique commands** with **8.3 hours** of recorded data. Furthermore, we augment the dataset with **10** commonly seen background sounds to simulate real scenarios, resulting in more than **80 hours** of data. We evaluate the collected data with **two baseline models** and observe a clear performance drop on the augmented data. This could be an interesting future research direction to tackle.

Dataset Collection

- **Template collection**. We collect multiple command templates covering into four general categories: 1) *navigation*; 2) *music playing*; 3) *weather inquiry*; and 4) *others*
- **Experts annotation**. We hire two human experts to filter out command patterns with high similarities to increase the diversity.
- **Template sampling**. To further increase the diversity of the generated commands, we uniformly sample ∼30% commands from the first three categories *navigation*, *music playing*, and *weather inquiry*) while keeping all the commands from the *others* category.
- **Resulting templates** We end up with 200 in-car commands of which 160 are from *navigation*, *music playing*, and *weather inquiry* categories and 40 are from the *others* category.

CI-AVSR Dataset

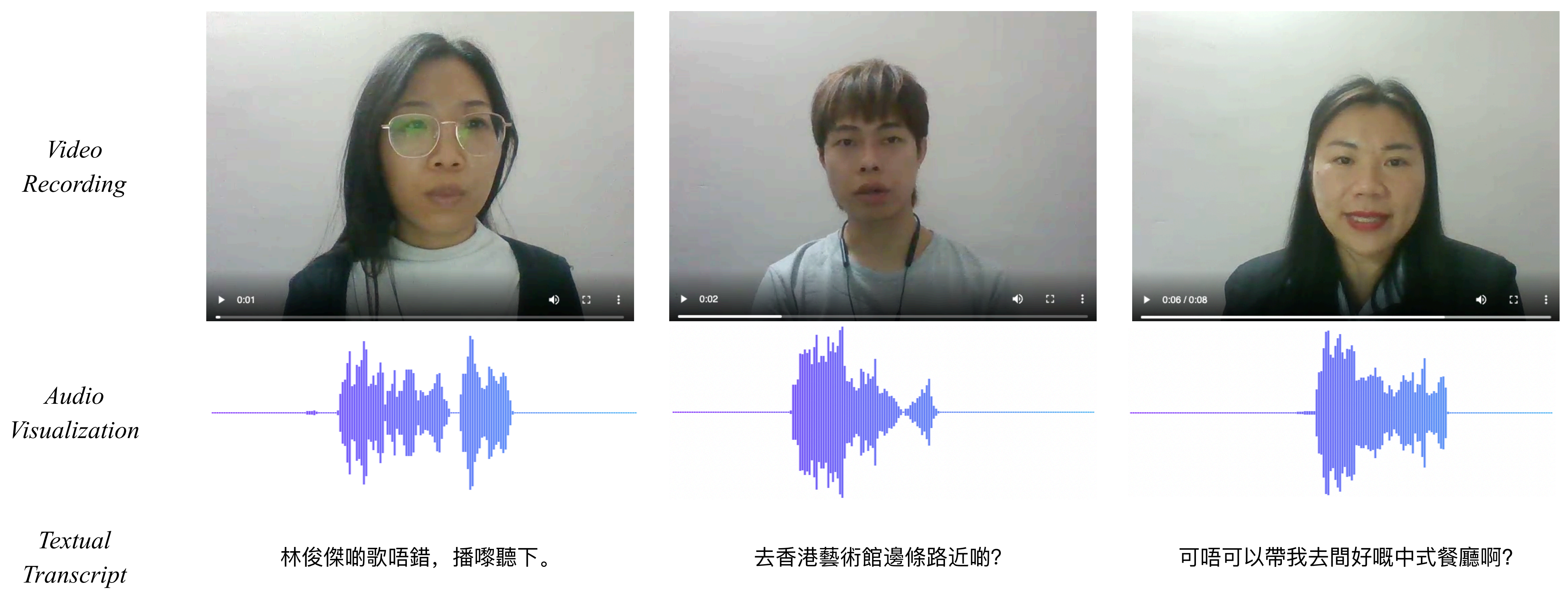

Figure 1:Examples of the CI-AVSR dataset. Each sample in the CI-AVSR consists of an audio-visual clip and the corresponding transcript. The task is to generate the transcript given the audio-visual information.

Category	Command Patterns	Complete Commands
Navigation	1. 導航唔該車我去[LOCATION]。 (Please navigate me to [LOCATION], thanks.) 2. 帶我去[LOCATION]呀，導航。 (Take me to [LOCATION], navigation.) 3. [LOCATION]，行邊條路最快去到？ ([LOCATION], what is the fastest path to go there?) 4. 邊條路去[LOCATION]最近？ (What is the shortest path to [LOCATION]?) 5. [LOCATION]可唔可以去到？ (Could you drive me to [LOCATION]?)	1. 導航唔該車我去香港科技大學。 (Please navigate me to HKUST, thanks.) 2. 導航唔該車我去香港藝術館。 (Please navigate me to the HK art museum, thanks.) 3. 帶我去尖沙咀呀，導航。 (Take me to Tsim Sha Tsui, navigation.) 4. 海洋公園，行邊條路最快去到？ (The ocean park, what is the fastest path to go there?) 5. 維多利亞港可唔可以去到？ (Could you drive me to the Victoria Park?)
Music Playing	1. 播放[SINGER]的[SONG]。 (Play [SINGER]'s [SONG].) 2. 我想聽[SONG]。 (I'd like to listen [SONG].) 3. 我想聽[SINGER]唱嘅歌。 (I want to hear a song by [SINGER].) 4. 嚟首[SINGER]的[SONG]。 (Play the [SONG] by [SINGER].) 5. [SINGER]歌唔錯，播嚟聽下。 (The songs by [SINGER] are good, play some please.)	1. 播放張國榮的我。 (Play 張國榮's 我.) 2. 我想聽海闊天空 (I'd like to listen 海闊天空.) 3. 我想聽陳奕迅唱嘅歌。 (I want to hear a song by 陳奕迅.) 4. 嚟首陳小春的亂世巨星。 (Play the [SONG] by [SINGER].) 5. 李克勤歌唔錯，播嚟聽下。 (The songs by [SINGER] are good, play some please.)
Weather Inquiry	1. [TIME]天氣如何？ (What's the forecast for [TIME]?) 2. 想睇下[TIME]嘅天氣點。 (I'd like to know the weather on [TIME].) 3. 幫我查下[TIME]嘅天氣。 (Please help me to check the weather [TIME].) 4. 唔該講下[TIME]天氣點啊？ (What's the weather [TIME]? Thanks.) 5. [TIME]天氣好，定係[TIME]天氣好？ (The weather [TIME] seems to be good, is it?)	1. 明天天氣如何？ (What's the forecast for tomorrow?) 2. 今天晚上天氣如何？ (What's the forecast for tonight?) 3. 唔該講下星期三天氣點啊？ (What's the weather on Wednesday? Thanks.) 4. 幫我查下禮拜天嘅天氣。 (Please help me to check the weather on Sunday.) 5. 週六天氣好，定係週六天氣好？ (The weather on Saturday seems to be good, is it?)

Table 3:Examples of command patterns, named entities, and the combination of them to form complete commands for the three categories, including *navigation*, *music playing*, and *weather inquiry*. English translations are provided in the parentheses, except the singers and songs that cannot be translated directly.

Split	#Male (Dur.)	#Female (Dur.)	Total Dur.
Train	10 (10,803s)	10 (11,813s)	22,616s
Valid	2 (1,849s)	2 (1,829s)	3,678s
Test	3 (1,902s)	3 (1,843s)	3,745s

Table 4:Statistics of the dataset split in CI-AVSR.

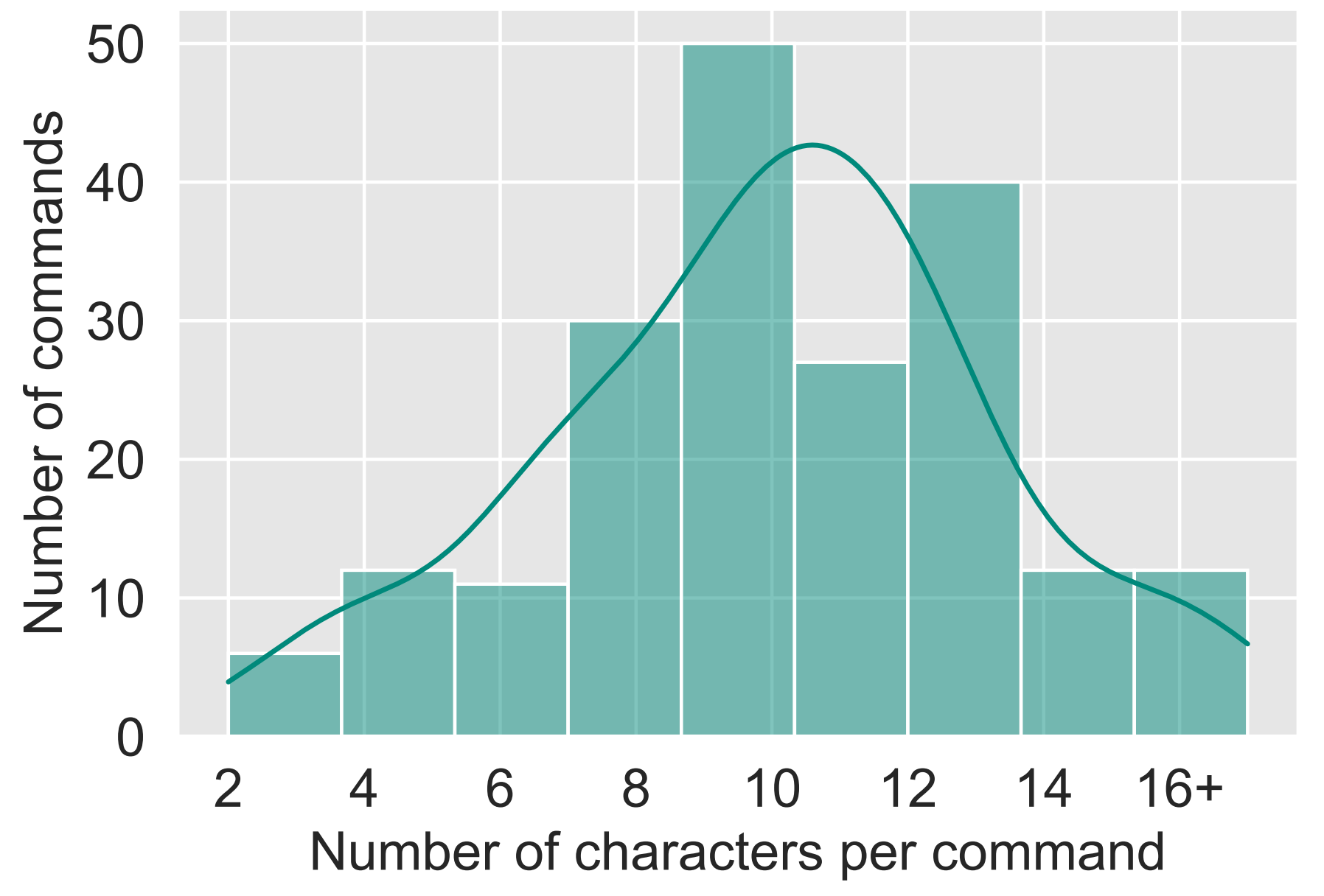

Figure 2:Number of characters distribution of all in-car commands in CI-AVSR.

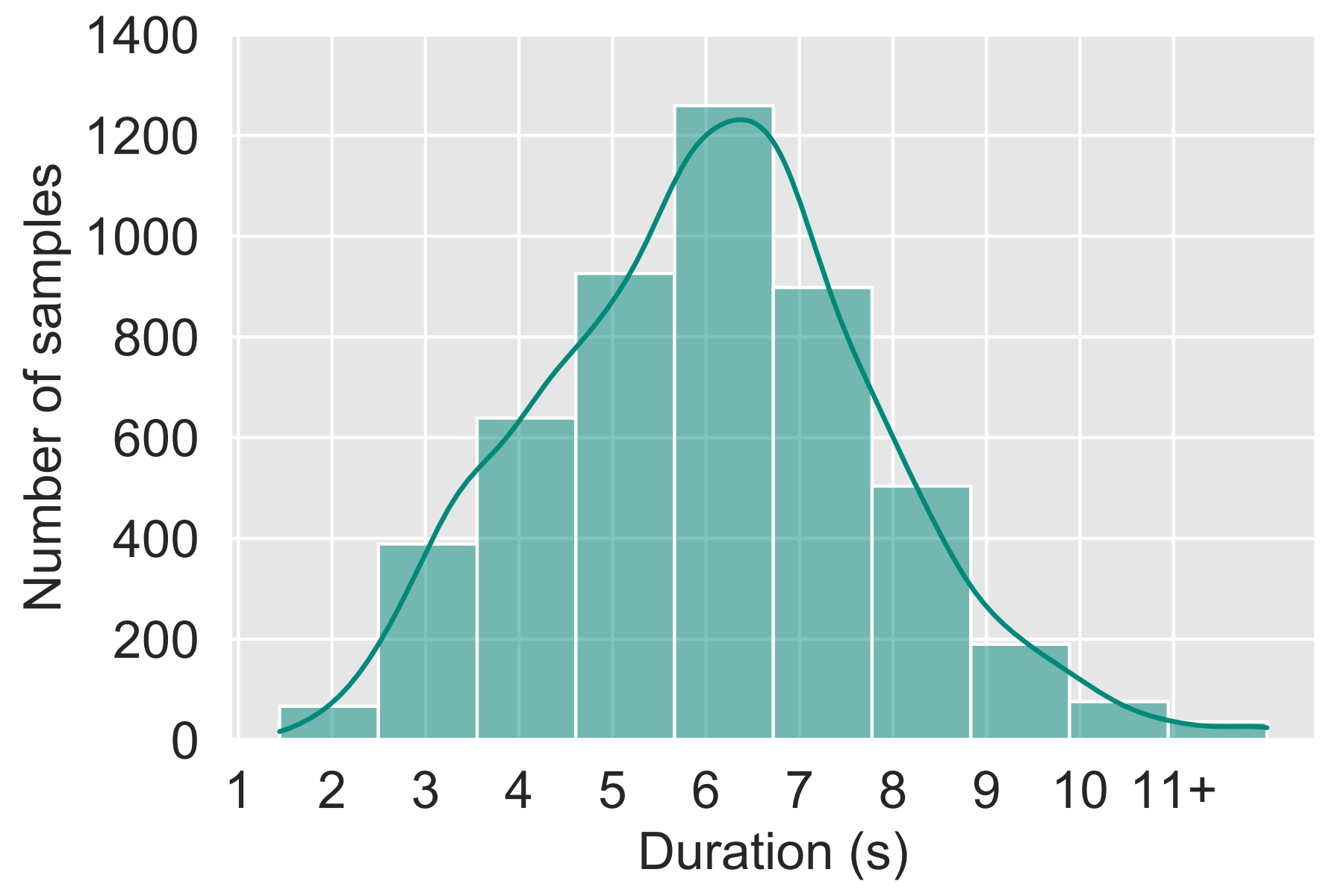

Figure 3:Duration distribution of all in-car commands in CI-AVSR.