Audio-Visual Understanding of Passenger Intents
for In-Cabin Conversational Agents

Eda Okur   **Shachi H Kumar**   **Saurav Sahay**   **Lama Nachman**

Intel Labs, Anticipatory Computing Lab, USA

{eda.okur, shachi.h.kumar, saurav.sahay, lama.nachman}@intel.com

Abstract

Building multimodal dialogue understanding capabilities situated in the in-cabin context is crucial to enhance passenger comfort in autonomous vehicle (AV) interaction systems. To this end, understanding passenger intents from spoken interactions and vehicle vision systems is an important building block for developing contextual and visually grounded conversational agents for AV. Towards this goal, we explore AMIE (Automated-vehicle Multimodal In-cabin Experience), the in-cabin agent responsible for handling multimodal passenger-vehicle interactions. In this work, we discuss the benefits of multimodal understanding of in-cabin utterances by incorporating verbal/language input together with the non-verbal/acoustic and visual input from inside and outside the vehicle. Our experimental results outperformed text-only baselines as we achieved improved performances for intent detection with multimodal approach.

1 Introduction

Understanding passenger intents from spoken interactions and visual cues (both from inside and outside the vehicle) is an important building block towards developing contextual and scene-aware dialogue systems for autonomous vehicles. When the passengers give instructions to the in-cabin agent AMIE, the agent should parse commands properly considering three modalities (i.e., verbal/language/text, vocal/audio, visual/video) and trigger the appropriate functionality of the AV system.

For in-cabin dialogue between car assistants and driver/passengers, recent studies explore creating a public dataset using a WoZ approach (Eric et al., 2017) and improving ASR for passenger speech recognition (Fukui et al., 2018). Another recent work (Zheng et al., 2017) attempts to classify sentences as navigation-related or not using the CU-Move in-vehicle speech corpus (Hansen et al., 2001), a relatively old and large corpus focusing on route navigation.

We collected a multimodal in-cabin dataset with multi-turn dialogues between the passengers and AMIE using a Wizard-of-Oz (WoZ) scheme via realistic scavenger hunt game. In previous work (Okur et al., 2019), we experimented with various RNN-based models to detect the utterance-level intents (i.e., *set-destination*, *change-route*, *go-faster*, *go-slower*, *stop*, *park*, *pull-over*, *drop-off*, *open-door*, *other*) along with the intent keywords and relevant slots (i.e., *location*, *position/direction*, *object*, *gesture/gaze*, *time-guidance*, *person*) associated with these intents.

In this work, we discuss the benefits of a multimodal understanding of in-cabin utterances by incorporating verbal/language input together with the non-verbal/acoustic and visual cues, both from inside and outside the vehicle (e.g., passenger gestures and gaze from in-cabin video stream, referred objects outside of the vehicle from the road view camera stream).

2 Data

Our AMIE in-cabin dataset includes 30 hours of multimodal data collected from 30 passengers (15 female, 15 male) in a total of 20 sessions. In 10 sessions, a single passenger was present, whereas the remaining 10 sessions include two passengers interacting with the vehicle. Participants sit in the back of the vehicle, separated from the driver and the human acting as an agent at the front. The vehicle is modified to hide the operator and the WoZ AMIE agent from the passengers, using a variation of the WoZ approach (Wang et al., 2017). In each ride/session, which lasted about 1 hour or more, the participants were playing a realistic scavenger hunt game on the streets of Richmond, BC,
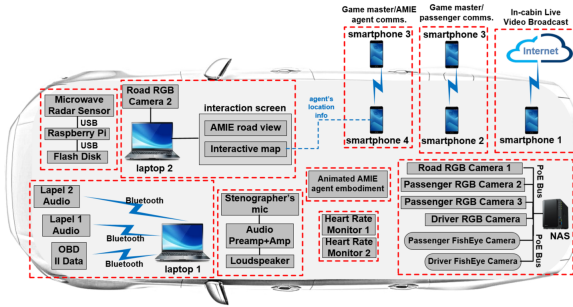
Figure 1: AMIE In-cabin Data Collection Setup

Canada. Passengers treat the vehicle as AV and communicate with the WoZ AMIE agent mainly via speech commands. Game objectives require passengers to interact naturally with the agent to go to certain destinations, update routes, give specific directions regarding where to pull over or park (sometimes with gestures), find landmarks (refer to outside objects), stop the vehicle, change speed, get in and out of the vehicle, etc. Further details of the data collection protocol and dataset statistics can be found in (Sherry et al., 2018; Okur et al., 2019). See Fig. 1 for the vehicle instrumentation to enable multimodal data collection setup.

2.1 Dataset Statistics

Multimodal AMIE dataset consists of in-cabin conversations between the passengers and the AV agent, with 10590 utterances in total. 1331 of these utterances have commands to the WoZ agent, hence they are associated with passenger intents. Utterance-level intent and word-level slot annotations are obtained on the transcribed utterances by majority voting of 3 annotators. The annotation results for *utterance-level intent* types, *slots* and *intent keywords* can be found in Table 1 and Table 2.

3 Methodology

We explored leveraging multimodality for the Natural Language Understanding (NLU) module in the Spoken Dialogue System (SDS) pipeline. As our AMIE in-cabin dataset has audio and video recordings, we investigated three modalities for the NLU: text, audio, and visual.

For text (verbal/language) modality, we employed the Hierarchical & Joint Bi-LSTM model (Schuster and Paliwal, 1997; Hakkani-Tur et al., 2016; Zhang and Wang, 2016; Wen et al., 2018), namely H-Joint-2.

- **Hierarchical & Joint Model (H-Joint-2)**: This is a 2-level hierarchical joint learning model that detects/extracts *intent keywords* & *slots* using sequence-to-sequence Bi-LSTMs first (Level-1), then only the words that are predicted as *intent keywords* & *valid slots* are fed into the Joint-2 model (Level-2), which is another sequence-to-sequence Bi-LSTM network for *utterance-level intent* detection, jointly trained with *slots* & *intent keywords*.

This architecture was chosen based on the best-performing uni-modal results presented in previous work (Okur et al., 2019) for utterance-level intent recognition and slot filling on our AMIE dataset. These initial uni-modal results were obtained on the transcribed text with pre-trained GloVe word embeddings (Pennington et al., 2014).

In this study, we explore the following multimodal features to better assess passenger intents for conversational agents in self-driving cars: word embeddings for text, speech embeddings and acoustic features for audio, and visual features for the video modality.

AMIE Scenario	Intent Type	Utterance Count
Set/Change Destination/Route	SetDestination	311
	SetRoute	507
	Park	151
Finishing the Trip	PullOver	34
	Stop	27
Set/Change Driving Behavior/Speed	GoFaster	73
	GoSlower	41
Others (Door, Music, A/C, etc.)	OpenDoor	136
	Other	51
	Total	*1331*

Table 1: AMIE In-cabin Dataset Statistics: Intents

Slot/Keyword Type	Word Count
Intent Keyword	2007
Location	1969
Position/Direction	1131
Person	404
Time Guidance	246
Gesture/Gaze	167
Object	110
None	6512
Total	*12546*

Table 2: AMIE In-cabin Dataset Statistics: Slots

Modalities	Features	F1(%)
Text	Word2Vec (37.6K vocab)	85.63
Text	GloVe (400K vocab)	89.02
Text & Audio	GloVe & Acoustic (openSMILE/IS10)	89.53
Text & Visual	GloVe & Video_cabin (CNN/Inception-ResNet)	89.40
Text & Visual	GloVe & Video_road (CNN/Inception-ResNet)	89.37
Text & Visual	GloVe & Video_cabin+road (CNN/Inception-ResNet)	89.68
Audio	Speech2Vec (37.6K vocab)	84.47
Text & Audio	Word2Vec+Speech2Vec	88.08
Text & Audio	GloVe+Speech2Vec	90.85
Text & Audio	GloVe+Word2Vec+Speech2Vec	91.29
Text & Audio	GloVe+Word2Vec+Speech2Vec & Acoustic (IS10)	91.68
Text & Audio & Visual	GloVe+Word2Vec+Speech2Vec & Video_cabin (CNN)	91.50
Text & Audio & Visual	GloVe+Word2Vec+Speech2Vec & Video_cabin+road (CNN)	91.55

Table 3: F1-scores of Intent Recognition with Multimodal Features

3.1 Word and Speech Embeddings

We incorporated pre-trained speech embeddings, called Speech2Vec[1], as additional audio-related features. These Speech2Vec embeddings (Chung and Glass, 2018) are trained on a corpus of 500 hours of speech from LibriSpeech. Speech2Vec can be considered as a speech version of Word2Vec embeddings (Mikolov et al., 2013), where the idea is that learning the representations directly from speech can capture the information carried by speech that may not exist in plain text.

We experimented with concatenating word and speech vectors using GloVe embeddings (6B tokens, 400K vocab, 100-dim), Speech2Vec embeddings (37.6K vocab, 100-dim), and its Word2Vec (37.6K vocab, 100-dim) counterpart, in which the Word2Vec embeddings are trained on the transcript of the same LibriSpeech corpus.

3.2 Acoustic Features

Using openSMILE[2] audio feature extraction toolkit (Eyben et al., 2013), 1582 acoustic features are extracted for each utterance using the segmented audio clips from AMIE dataset. These are the INTERSPEECH 2010 Paralinguistic Challenge (IS10) features (Schuller et al., 2010) including PCM (pulse-code modulation) loudness, MFCC (Mel-frequency cepstral coefficients), log Mel Freq. Band, LSP (line spectral pairs) Frequency, etc.

3.3 Visual Features

Intermediate CNN features[3] are extracted from each video clip segmented per utterance from the AMIE dataset. Using the feature extraction process described in (Kordopatis-Zilos et al., 2017), one frame per second is sampled for any given input video clip and its visual descriptors are extracted from the activations of the intermediate convolution layers of a pre-trained CNN. We used the pre-trained Inception-ResNet-v2 model[4] (Szegedy et al., 2016) and generated 4096-dim features for each sample. We experimented with utilizing two sources of visual information: (i) cabin/passenger view from the back-driver RGB camera recordings, (ii) road/outside view from the dash-cam RGB video streams.

4 Experimental Results

Performance results of the utterance-level intent recognition models with varying modality and feature concatenations can be found in Table 3, using hierarchical joint learning (H-Joint-2). For text and speech embeddings experiments, we observe that using Word2Vec or Speech2Vec representations achieve comparable F1-score performances, which are significantly below the GloVe embeddings performance. This was expected as the pre-trained Speech2Vec vectors have lower vocabulary coverage than the GloVe vectors. On the

other hand, we observe that concatenating GloVe + Speech2Vec embeddings, and further GloVe + Word2Vec + Speech2Vec yields higher F1-scores for intent recognition. These results show that the speech embeddings indeed can capture useful semantic information carried by speech only, which may not exist in plain text.

We also investigate incorporating the audio-visual features on top of text-only and text + speech embedding models. Including openSMILE/IS10 acoustic features from audio as well as intermediate CNN/Inception-ResNet-v2 features from video brings slight improvements to our intent recognition models, achieving 0.92 F1-score. These initial results may require further explorations for specific intents such as *stop* (e.g., audio intensity & loudness could have helped), or for relevant slots such as passenger *gesture/gaze* (e.g., cabin-view features) and outside *object*s (e.g., road-view features).

5 Conclusion and Future Work

In this work, we briefly present our initial explorations towards the multimodal understanding of passenger utterances in autonomous vehicles. We show that our experimental results outperformed the uni-modal text-only baseline results, and with multimodality, we achieved improved performances for passenger intent detection in AV. This ongoing research has the potential impact of exploring real-world challenges with human-vehicle-scene interactions for autonomous driving support via spoken utterances.

There exist various exciting recent work on improved multimodal fusion techniques (Zadeh et al., 2018; Liang et al., 2019a; Pham et al., 2019; Baltrušaitis et al., 2019). In addition to the simplified feature and modality concatenations, we plan to explore some of these promising tensor-based multimodal fusion networks (Liu et al., 2018; Liang et al., 2019b; Tsai et al., 2019) for more robust intent classification on AMIE dataset as future work.

References

T. Baltrušaitis, C. Ahuja, and L. Morency. 2019. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443.

Yu-An Chung and James Glass. 2018. Speech2vec: A sequence-to-sequence framework for learning word embeddings from speech. In *Proc. INTERSPEECH 2018*, pages 811–815.

Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D. Manning. 2017. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 37–49. Association for Computational Linguistics.

Florian Eyben, Felix Weninger, Florian Gross, and Björn Schuller. 2013. Recent developments in opensmile, the munich open-source multimedia feature extractor. In *Proc. ACM International Conference on Multimedia*, MM '13, pages 835–838.

M. Fukui, T. Watanabe, and M. Kanazawa. 2018. Sound source separation for plural passenger speech recognition in smart mobility system. *IEEE Transactions on Consumer Electronics*, 64(3):399–405.

Dilek Hakkani-Tur, Gokhan Tur, Asli Celikyilmaz, Yun-Nung Vivian Chen, Jianfeng Gao, Li Deng, and Ye-Yi Wang. 2016. Multi-domain joint semantic frame parsing using bi-directional rnn-lstm. ISCA.

John HL Hansen, Pongtep Angkititrakul, Jay Plucienkowski, Stephen Gallant, Umit Yapanel, Bryan Pellom, Wayne Ward, and Ron Cole. 2001. Cu-move: Analysis & corpus development for interactive in-vehicle speech systems. In *Seventh European Conference on Speech Communication and Technology*.

Giorgos Kordopatis-Zilos, Symeon Papadopoulos, Ioannis Patras, and Yiannis Kompatsiaris. 2017. Near-duplicate video retrieval by aggregating intermediate cnn layers. In *International Conference on Multimedia Modeling*, pages 251–263. Springer.

Paul Pu Liang, Yao Chong Lim, Yao-Hung Hubert Tsai, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2019a. Strong and simple baselines for multimodal utterance embeddings. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2599–2609, Minneapolis, Minnesota. Association for Computational Linguistics.

Paul Pu Liang, Zhun Liu, Yao-Hung Hubert Tsai, Qibin Zhao, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2019b. Learning representations from imperfect time series data via tensor rank regularization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1569–1576, Florence, Italy. Association for Computational Linguistics.

Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, USA.

Eda Okur, Shachi H Kumar, Saurav Sahay, Asli Arslan Esme, and Lama Nachman. 2019. Natural language interactions in autonomous vehicles: Intent detection and slot filling from passenger utterances. *20th International Conference on Computational Linguistics and Intelligent Text Processing (CICLing 2019)*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP'14)*.

Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6892–6899.

Björn Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian Müller, and Shrikanth S Narayanan. 2010. The interspeech 2010 paralinguistic challenge. In *Proc. INTERSPEECH 2010*.

M. Schuster and K.K. Paliwal. 1997. Bidirectional recurrent neural networks. *Trans. Sig. Proc.*, 45(11):2673–2681.

John Sherry, Richard Beckwith, Asli Arslan Esme, and Cagri Tanriover. 2018. Getting things done in an autonomous vehicle. In *Social Robots in the Wild Workshop, 13th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2018)*.

Christian Szegedy, Sergey Ioffe, and Vincent Vanhoucke. 2016. Inception-v4, inception-resnet and the impact of residual connections on learning. *CoRR*, abs/1602.07261.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Florence, Italy. Association for Computational Linguistics.

Peter Wang, Srinath Sibi, Brian Mok, and Wendy Ju. 2017. Marionette: Enabling on-road wizard-of-oz autonomous driving studies. In *Proceedings of the 2017 ACM/IEEE International Conference on Human-Robot Interaction*, HRI '17, pages 234–243, New York, NY, USA. ACM.

Liyun Wen, Xiaojie Wang, Zhenjiang Dong, and Hong Chen. 2018. Jointly modeling intent identification and slot filling with contextual and hierarchical information. In *Natural Language Processing and Chinese Computing*, pages 3–15, Cham. Springer.

Amir Zadeh, Paul Pu Liang, Navonil Mazumder, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Memory fusion network for multi-view sequential learning. *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*.

Xiaodong Zhang and Houfeng Wang. 2016. A joint model of intent determination and slot filling for spoken language understanding. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, IJCAI'16, pages 2993–2999.

Y. Zheng, Y. Liu, and J. H. L. Hansen. 2017. Navigation-orientated natural spoken language understanding for intelligent vehicle dialogue. In *2017 IEEE Intelligent Vehicles Symposium (IV)*, pages 559–564.