A Multilingual Speech-Based Driver Assistant for Basque and English

Antonio Aparicio Akcharov[1], **Asier López Zorrilla**[1], **Juan Camilo Vasquez-Correa**[2], **Oscar Montserrat**[3], **Jose María Echevarría**[3], **Begoña Arrate**[4], **Joxean Zapirain**[4], **Mikel deVelasco**[1], **Santiago Andrés Moreno-Acevedo**[2], **Ander González-Docasal**[2], **M. Inés Torres**[1], and **Aitor Álvarez**[2]

[1]University of the Basque Country UPV/EHU, [2]Vicomtech
[3]Automotive Intelligence Centre AIC, [4]UZEI

Abstract

This demo paper presents a prototype of a multilingual, speech-based driver assistant, designed to support both English and Basque languages. The inclusion of Basque—a low-resource language with limited domain-specific training data—marks a significant contribution, as publicly available AI models, including Large Language Models, often underperform for such languages compared to high-resource languages like English. Despite these challenges, our system demonstrates robust performance, successfully understanding user queries and delivering rapid responses in a demanding environment: a car simulator. Notably, the system achieves comparable performance in both English and Basque, showcasing its effectiveness in addressing linguistic disparities in AI-driven applications. A demo of our prototype will be available in the workshop.

1 Introduction

Speech-based driver assistants have become increasingly prevalent in modern vehicles, offering convenience and safety by enabling hands-free interaction with in-car systems. Prominent examples such as Amazon Alexa Auto or Apple CarPlay, have demonstrated the potential of such technologies to enhance user experience by providing real-time information about navigation, vehicle status, and other essential tasks (Li et al., 2024; Zhou and Zheng, 2023). These systems leverage advancements in Natural Language Processing (NLP) and speech recognition, often relying on Large Language Models (LLMs) trained on extensive datasets mainly containing data in high-resource languages like English (Touvron et al., 2023; Günther et al., 2023). However, due to the scarcity of domain-specific datasets and models, the development of similar systems for low-resource languages remains a significant challenge.

Low-resource languages often lack the domain-specific annotated data necessary to train state-of-the-art NLP and speech processing models, leading to a greater reliance on rule-based approaches (Wilcock et al., 2017). Basque is an isolated language spoken by approximately 750,000 people, with complex morphology and free word order. Despite recent progress, publicly available Basque language AI models such as the Latxa series (Etxaniz et al., 2024) or Llama-eus (Corral et al., 2024) still underperform compared to models for high-resource languages. Additionally, there are limited datasets and tools tailored for specific applications, such as automotive environments, further complicating the development of Basque-focused driver assistants. In fact, we have had to develop an automotive industry-related dataset from scratch for this project, since most publicly available and related datasets are in English or high-resource languages (Deruyttere et al., 2020).

This demo paper presents a multilingual speech-based driver assistant developed within the Adapt-IA project, designed to provide support in both Basque and English. Integrated into an industrial car simulator, the assistant provides conversational access to real-time vehicle data, including speed, traffic conditions, tire pressure, and battery status. By creating a domain-specific corpus in Basque and leveraging state-of-the-art NLP and speech processing techniques, our system, which will be shown in a demo in the workshop, demonstrates comparable performance in Basque and English. This showcases the feasibility of extending driver assistant technology to underrepresented languages like Basque.

The rest of the paper is organised as fol-

Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology, pages 190–195
May 27–30, 2025.

lows: Section 2 provides an overview of the context and motivation behind this work. Section 3 describes the corpus built to develop our assistant, and its architecture and implementation. Section 4 presents a sample dialogue that showcases its capabilities. Finally, Section 5 presents our conclusions.

2 Context

This work has been carried out within the Adapt-IA project, which aims to develop AI technologies for the Basque language, and others if the specific use case so requires. More particularly, the primary goal is to explore the development and integration of these technologies into the needs and specific use cases of various Basque industrial sectors, such as machine-tools manufacturing, the energy sector, the railway sector and the automotive industry.

One sector where these advancements have significant potential is the automotive industry. As vehicles become increasingly reliant on intelligent systems, the need for speech-based technologies in different languages, including Basque, becomes critical. In this context, this work represents a step towards providing drivers with the capabilities to naturally interact with smart vehicles in Basque and English languages. To this end, we have developed a driving assistant that suits the Automotive Intelligence Center (AIC) industrial car simulator in the Basque country: an hexapod system equipped with a fully operational cockpit interior from a road vehicle (see Figure 1). With six degrees of freedom, this simulator is employed for dynamic analyses and serves as a platform for testing advanced driver-assistance systems and user communication strategies under highly realistic conditions.

3 The Adapt-IA Speech-Based Driver Assistant

This section describes all the stages and tools created and integrated into the developed speech-based driver assistant, including the data generation process, in addition to the training of specific models for automatic speech recognition (ASR), natural language understading (NLU), and text to speech (TTS) that have been integrated into the final prototype.

Figure 1: A user talking to the Adapt-IA driver assistant in the AIC simulator.

3.1 Data generation

Collecting training data is a critical initial phase in the development of any AI system. In this case, one of our main objectives was to create a specialised dataset with specific terminology about the automotive industry. We built a corpus comprising over 16K in-car interaction scenarios in both Basque and English. covering a set of 14 intents (see Table 1). The full dataset was filtered using different strategies and used to train both the NLU and ASR systems of the developed assistant.

We tackled the data generation task through two separate strategies for each language.

English

Recent advancements in LLMs enabled us to efficiently generate artificial data for the English use case. Initially, we curated a small organic dataset comprising example queries that a user might pose to their car assistant, such as "How many kilometers of autonomy do I have left?". These examples served as the foundation for generating additional data.

Using state-of-the-art LLMs (ChatGPT-4o mainly), we expanded the dataset by synthesizing hundreds of sentences based on the original examples. This yielded a dataset of 1 700 unique sentences. The generated data was thoroughly evaluated and found to meet the requirements of our use case, proving the effectiveness of LLMs for synthetic data generation.

request speed
request speed limit
request tire pressure
request battery level
request autonomy
request driving time
request driving distance
request traffic status
repeat
car did not understand
thanks
hello
goodbye
other

Table 1: List of user intents the assistant is able to understand. The "car did not understand" label is used when user says that the car did not provide the expected information, and the "other" indicates that the information the user is requesting is not available for this system.

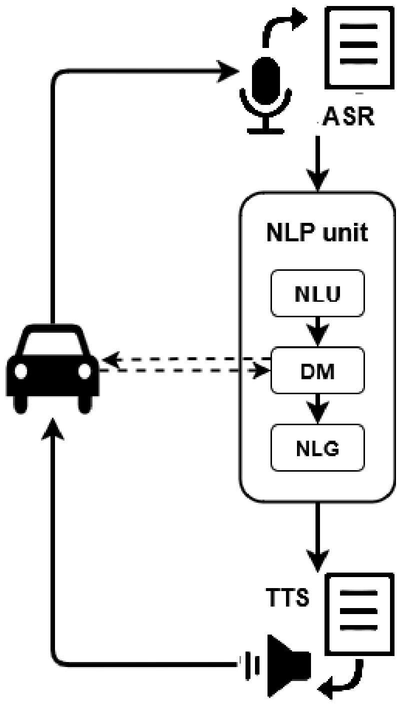

Figure 2: High-level overview of the assistant's back end architecture.

Basque

One of the main goals of the project was to generate high-quality data in Basque with specialised terminology for different industrial sectors, and which could be useful for the community and for future works. Therefore, the use of LLMs was not desired in this case, as their performance remains limited for low-resource languages (Hasan et al., 2024; Jayakody and Dias, 2024).

Instead, we built a dataset about how driving information could be asked in Basque. Expert Basque linguists generated a number of templates to form sentences, which considered different word orders within sentences, synonyms, possible word omissions, and different registers (formal, informal, direct, etc.). More than 20 million sentences were generated considering all possible combinations. However, we filtered very similar sentences to make the corpus as diverse as possible. To this end, we employed both a cluster-based active learning approach (Moreno-Acevedo et al., 2024), and a random selection following a uniform distribution. This led to a diverse and balanced corpus of around 14K sentences.

Additionally, we asked some volunteers how they would form sentences for the defined user-intents. This way, we increased the dataset by another 612 natural sentences[1].

3.2 System architecture and components

As for the assistant interface, the user interacts with it via a tablet mounted on the panel of the car simulator, which acts both as the microphone and the speaker. If the user wants to ask anything while driving, they can press a button and start a conversation. This conversation ends when the user presses the button again, or when the system determines that the dialogue has ended (if the user says "bye", for example).

Since some modules require relatively intensive computing capabilities, the processing is done on remote servers. There are three main components, as depicted in Figure 2. The communication between these modules is done via API queries. It takes around 3 seconds to process the user's speech and generate a speech response. The main specifications for each component are described as follows.

ASR

The ASR unit is built on top of Nvidia's Parakeet models, which are state-of-the-art recurrent neural network transducers (RNN-T) that have shown excellent results in recent speech

[1]This corpus will be publicly available once the project is finished.

benchmarks (Srivastav et al., 2025; Vásquez-Correa et al., 2024). The Basque and English models use the large version of the fast-conformer RNN-T architecture, which has 0.6 billion parameters[2]. The Basque model was initially trained from scratch using 1 258 hours of transcribed Basque audio and then fine-tuned with 12.8 hours of in-domain data. This in-domain data was created by synthesizing 10 166 selected user requests using our TTS system. The English model started with Nvidia's pre-trained model and was fine-tuned using 7 hours of in-domain data, also created by synthesizing 8 223 user requests using our TTS system. In this case, in addition to the English corpus described in Section 3.1, we automatically translated the filtered Basque sentences into English to increase the amount of data.

NLP unit

The NLP unit consists of three modules:

1. NLU: Takes a transcribed user sentence and outputs the corresponding intents; which represents the semantic meaning of the sentence.
2. DM (Dialogue Manager): Based on the user intents and the dialogue history, the DM outputs the next dialogue act; i.e., what the system needs to reply.
3. NLG (Natural Language Generation): Generates a natural sentence based on dialogue act produced by the DM.

Due to the limitations of Basque NLP models, we employed different strategies for the NLU. In English, the user request is transformed into a 384-dimensional embedding vector with the model `all-MiniLM-L12-v2` from Sentence-Transformers (Wang et al., 2020). Conversely, for the Basque case, we had to use an LLM fine-tuned specifically for this language. In this case we pass the sentence through Latxa 7B (Etxaniz et al., 2024) and retrieve the output embedding. In both cases, we train multilayer neural networks to classify the embeddings into one or more user intents. Those are trained on the datasets we gathered. The models are run in parallel, so that switching between languages is effortless.

[2] https://huggingface.co/nvidia/parakeet-rnnt-0.6b

Once the intents of the input are obtained, the rule-based DM retrieves the relevant information from the car simulator, outputs the next dialogue act. Generally, these rules involve providing the requested information, checking the dialogue history if a repetition of information is requested, and greeting or thanking the user when appropriate. One key advantage of rule-based DMs is their greater controllability and robustness, particularly in restricted environments (Vázquez et al., 2023).

Given the dialogue act output by the DM, the template-based NLG module generates a textual answer to the user's query. To avoid repetitions in the assistant's responses, we have a number (around 10) templates per dialogue act. These are selected randomly and filled with the real-time data provided by the simulator.

TTS

The integrated TTS system was based on a Tacotron-2 (Shen et al., 2018) acoustic model. This architecture is built using an encoder-decoder with attention that returns the Mel-scale spectrogram of the generated speech, which is in turn converted to an audio signal by a HiFi-GAN (Kong et al., 2020) neural vocoder.

A Tacotron-2 model was trained per language on pairs of audio and their corresponding transcription. The models were fine-tuned from a pre-existing model trained on 34 137 audio files consisting of 36.1 h of high-quality studio recordings from a female Spanish speaker. The Basque dataset comprised 11 640 audio files summing a total of 20.8 h standard Basque recordings from a native speaker, whilst for English the public domain LJ-Speech dataset (Ito and Johnson, 2017) was used. The input string consisted of characters in the case of Basque and phonemes for English, which were generated using a DeepPhonemizer[3] model adapted from pronunciation data extracted from Wiktionary. Finally, the HiFi-GAN vocoder made use of the ground-truth aligned spectrograms from the same datasets needed for training the acoustic models.

[3] https://github.com/as-ideas/DeepPhonemizer

Basque	English
User: Kaixo kotxe, nire abiadura legala al da?	**User:** Hi there, is my current speed correct?
Assistant: Abiadura muga 90 kilometro ordukoa da, eta zure abiadura 87 kilometro ordukoa.	**Assistant:** The speed limit is 90 kilometers per hour, while your speed is 87 kilometers per hour.
User: Eta zenbateko autonomia daukagu?	**User:** And how far can we drive with the current battery?
Assistant: Bateriaren karga honekin, 250 kilometro gehiago egin ditzakegu.	**Assistant:** With this battery charge, we can drive 250 more kilometers.
User: Barka, ez zaitut ondo entzun...	**User:** Sorry, I couldn't hear you well...
Assistant: 250 kilometro egin ditzakegu.	**Assistant:** We can drive 250 more kilometers.
User: Zein da kanpoaldeko tenperatura?	**User:** What's the temperature outside?
Assistant: Barkatu, baina ez dut horretan laguntzeko gaitasunik.	**Assistant:** I'm afraid I cannot provide this information.
User: Ongi da. Eskerrik asko eta agur!	**User:** It's OK. Thanks! Bye!
Assistant: Ez horregatik. Agur!	**Assistant:** You are welcome. Goodbye!

Table 2: A (parallel) sample conversation with the Adapt-IA assistant, in Basque and English.

4 A sample conversation with the Adapt-IA Driver Assistant

Table 2 shows the kind of dialogues implemented in the Adapt-IA driver assistant in Basque and English. Notice that the system is able to provide the requested information, or inform that some information is not available to the system, such as the outside temperature. It can also handle repetition requests if the user did not understand what the assistant said for any reason, and react to task-free intents such as greetings or gratitude.

A demo of this prototype where all the capabilities of the system will be shown will be presented in the conference.

5 Conclusion

We have presented a speech-based driving assistant capable of providing relevant real-time data to the driver in a conversational fashion, both in Basque and English. To achieve this, we developed an organic domain-specific corpus in Basque to train our system, effectively bridging the performance gap between the English and Basque versions. This effort demonstrates the feasibility of extending similar NLP systems to low-resource languages, such as Basque, thereby addressing the linguistic imbalance often found in AI technologies. Our results demonstrate that tailored approaches can overcome resource limitations, enabling robust performance for underrepresented languages in specialized domains. This work not only advances the feasibility of developing Basque-centric AI systems but also contributes to the broader goal of inclusivity in AI technologies for diverse linguistic communities.

Acknowledgments

This work was partially supported by the Department of Economic Development and Competitiveness of the Basque Government (Spri Group) through funding for the ADAPT-IA project (KK-2023/00035).

References

Ander Corral, Ixak Sarasua, and Xabier Saralegi. 2024. Pipeline analysis for developing instruct llms in low-resource languages: A case study on basque. *Preprint*, arXiv:2412.13922.

Thierry Deruyttere, Simon Vandenhende, Dusan Grujicic, Yu Liu, Luc Van Gool, Matthew Blaschko, Tinne Tuytelaars, and Marie-Francine Moens. 2020. Commands 4 autonomous vehicles (c4av) workshop summary. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 3–26. Springer.

Julen Etxaniz, Oscar Sainz, Naiara Perez, Itziar Aldabe, German Rigau, Eneko Agirre, Aitor Ormazabal, Mikel Artetxe, and Aitor Soroa. 2024. Latxa: An open language model and evaluation suite for basque. *Preprint*, arXiv:2403.20266.

Michael Günther, Georgios Mastrapas, Bo Wang, Han Xiao, and Jonathan Geuter. 2023. Jina embeddings: A novel set of high-performance sentence embedding models. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 8–18.

Md Arid Hasan, Prerona Tarannum, Krishno Dey, Imran Razzak, and Usman Naseem. 2024. Do large language models speak all languages equally? a comparative study in low-resource settings. *arXiv preprint arXiv:2408.02237*.

Keith Ito and Linda Johnson. 2017. The lj speech dataset. https://keithito.com/LJ-Speech-Dataset/.

Ravindu Jayakody and Gihan Dias. 2024. Performance of recent large language models for a low-resourced language. In *2024 International Conference on Asian Language Processing (IALP)*, pages 162–167. IEEE.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. 2020. HiFi-GAN: Generative adversarial networks for efficient and high fidelity speech synthesis. *Advances in neural information processing systems*, 33:17022–17033.

Yuanchao Li, Lachlan Urquhart, Nihan Karatas, Shun Shao, Hiroshi Ishiguro, and Xun Shen. 2024. Beyond voice assistants: Exploring advantages and risks of an in-car social robot in real driving scenarios. *arXiv preprint arXiv:2402.11853*.

Santiago A Moreno-Acevedo, Juan Camilo Vasquez-Correa, Juan M Martín-Doñas, and Aitor Álvarez. 2024. Stream-based active learning for speech emotion recognition via hybrid data selection and continuous learning. In *International Conference on Text, Speech, and Dialogue*, pages 105–117. Springer.

Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, Rif A. Saurous, Yannis Agiomvrgiannakis, and Yonghui Wu. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4779–4783.

Vaibhav Srivastav, Somshubra Majumdar, Nithin Koluguri, Adel Moumen, Sanchit Gandhi, et al. 2025. Open automatic speech recognition leaderboard. https://huggingface.co/spaces/hf-audio/open_asr_leaderboard.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Juan Camilo Vásquez-Correa, Aitor Álvarez, Haritz Arzelus, Santiago Andrés Moreno Acevedo, Ander González-Docasal, and Juan Manuel Martín-Doñas. 2024. The vicomtech speech transcription systems for the albayzín 2024 bilingual basque-spanish speech to text (bbs-s2t) challenge. In *Proc. IberSPEECH 2024*, pages 305–309.

Alain Vázquez, Asier López Zorrilla, Javier Mikel Olaso, and María Inés Torres. 2023. Dialogue management and language generation for a robust conversational virtual coach: Validation and user study. *Sensors*, 23(3):1423.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Preprint*, arXiv:2002.10957.

Graham Wilcock, Niklas Laxström, Juho Leinonen, Peter Smit, Mikko Kurimo, and Kristiina Jokinen. 2017. *Towards SamiTalk: A Sami-Speaking Robot Linked to Sami Wikipedia*, pages 343–351. Springer Singapore, Singapore.

Xinyi Zhou and Yangshuo Zheng. 2023. Research on personality traits of in-vehicle intelligent voice assistants to enhance driving experience. In *International Conference on Human-Computer Interaction*, pages 236–244. Springer.