

MentorIA: Developmental Cognitive Tutoring for Children through Specialized Speech Technologies*

Jaione Macicior-Mitxelena^{1,**}, Ricardo-Manuel García-Marreros¹, Unai Pérez-Goya¹ and José Javier Astrain¹

¹Department of Statistics, Computer Science and Mathematics. Public University of Navarre, 31006 Pamplona (Spain)

Abstract

This paper presents MentorIA, an interdisciplinary project aimed at developing an automated reading tutor for children in Basque-medium education. The system integrates Automatic Speech Recognition (ASR), error detection, and pedagogical modules to support literacy development in a limited-resourced language. Unlike conventional ASR systems that prioritise fluent and normalised outputs, this application requires strictly literal transcriptions to preserve diagnostic reading errors such as substitutions, omissions, and repetitions. We focus on adapting a Whisper-based autoregressive model to this setting, addressing its inherent tendency to smooth disfluencies. A multi-stage fine-tuning strategy combining adult speech and synthetic error-infused data is proposed. We observe that incorporating synthetic data improves the model's ability to retain non-canonical speech patterns and generalise to real children's speech. These findings highlight the importance of task-specific adaptation when deploying ASR within educational diagnostic systems.

Keywords

Automatic Speech Recognition, Basque Language, Children's Speech, Literal Transcription, Educational Technology

1. Introduction

Reading acquisition is a complex process that requires continuous guidance and individualised feedback from teachers. However, in real classroom contexts, teachers often face the challenge of simultaneously addressing the diverse needs of many students, which is particularly demanding in multilingual regions such as Navarre, where Basque is a low-resource language and many children enrolled are not native speakers. For these learners, mastering the language can be particularly difficult during early education (ages 6–10), and the situation is further complicated by the fact that their parents often lack sufficient knowledge of Basque to support them at home. Consequently, teachers must compensate for these limitations, making their task both time-consuming and complex. To address these challenges, this research is conducted within an educational project focused on designing an application capable of automatically analysing children's reading-aloud exercises. The application aims to assess reading performance, identify specific reading errors, and support learners through personalised feedback and adaptive educational games that offer targeted practice tailored to each child's difficulties.

To facilitate learning, the system must be capable of automatically identifying and classifying a wide range of reading difficulties and reporting them to teachers. These reports compile evidence of detected difficulties, providing teachers with a clear and actionable overview of each student's performance. In addition, a dashboard enables teachers to monitor students' learning progress both individually and collectively.

The central component is automatic speech recognition (ASR), which in this context must accurately

SEPLN 2026: 42nd International Conference of the Spanish Society for Natural Language Processing, León, Spain, 22-25 September 2026.

* Project Title: MentorIA: Desarrollo cognitivo infantil tutorizado (MENTORIA). Reference: 011-1365-2024-000121.

** Corresponding author.

✉ jaione.macicior@unavarra.es (J. Macicior-Mitxelena); ricardo.garcia@unavarra.es (R. García-Marreros); unai.perez@unavarra.es (U. Pérez-Goya); josej.astrain@unavarra.es (J.J. Astrain)

🆔 0009-0001-7392-4226 (J. Macicior-Mitxelena); 0000-0002-0729-1467 (R. García-Marreros); 0000-0002-2796-9079 (U. Pérez-Goya); 0000-0002-7792-6317 (J.J. Astrain)



© 2026 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

detect and characterise the errors produced by students during oral reading. In contrast to conventional ASR systems, the objective is not to infer the intended lexical item, but to generate a faithful transcription of the speech as it was actually produced. This makes it possible to identify inappropriate pauses, repetitions, and misreadings, all of which may provide evidence of reading difficulties. The resulting information can then be interpreted by specialists to inform targeted remediation and the design of personalized learning activities.

Within this broader framework, the present work focuses on the ASR component for the task of generating literal transcriptions in Basque. Special attention is given to two key aspects: the preservation of literal transcriptions and the robustness of the model to the acoustic variability of children’s speech. To this end, we analyse the model behaviour using different types of speech data, including adult, children’s, and synthetic speech. Each dataset provides complementary information about the model’s behaviour, allowing us to analyse its ability to adapt to Basque, produce faithful transcriptions, and handle the specific characteristics of children’s speech.

Overall, this work contributes to the development of reliable ASR technology within an educational application aimed at improving reading skills in a low-resource language context. From a research perspective, this work focuses on adapting an autoregressive ASR model to preserve literal transcriptions in children’s speech. In particular, we propose a multi-stage fine-tuning strategy that combines adult speech with synthetic error-infused data to mitigate the normalisation bias of standard ASR systems.

2. Related Work

In line with the framework described above, research into related work has focused on the evolution of ASR from Gaussian Mixture Model - Hidden Markov Model (GMM-HMM) pipelines to end-to-end neural models, such as Recurrent Neural Networks (RNNs) and Transformers, has significantly improved scalability and generalisation [1, 2]. Modern architectures leverage self-attention mechanisms to model long-range dependencies, outperforming traditional models in efficiency and representational power [3, 4]. High-performing models like *wav2vec 2.0* [5] and *Whisper* [6] have introduced self-supervised pretraining on massive datasets, while multilingual variants like *XLS-R* facilitate adaptation to over 128 languages [7]. However, adapting these large-scale models to low-resource languages and specialized domains remains a persistent challenge [8, 9].

Although hybrid DNN-HMM (Deep Neural Networks – Hidden Markov Models) [10] and GMM-HMM [11] systems represent stable and well-established approaches, they have been largely outperformed by end-to-end models across many benchmark tasks [12, 13, 14]. Self-supervised models such as *Wav2Vec 2.0* may exhibit degradation in noisy or acoustically mismatched conditions when compared to large-scale robust systems; however, they demonstrate strong cross-lingual and cross-accent generalisation and achieve competitive performance after pretraining. The *Whisper* model offers high robustness, particularly for noisy and conversational speech, albeit at a higher computational cost compared to smaller or optimised architectures. *WhisperX* [15] further improves performance by providing strong automated metrics and enhanced alignment for downstream tasks, although its effectiveness may vary depending on speech disfluencies and the behaviour of the Voice Activity Detection (VAD) component. In contrast, while Google ASR performs competitively on certain subtitle-related and disfluency-focused metrics, it offers limited transparency and controllability relative to open-source models. Finally, speech-to-text (STT) systems such as *Vosk* and *DeepSpeech* are often included as open-source reference baselines due to their lightweight deployment and historical relevance; nevertheless, they generally lag behind modern transformer-based models in terms of recognition accuracy.

In educational contexts, systems are traditionally evaluated using Word Error Rate (WER) [16]. However, conventional ASR models are typically optimized for semantic correctness and tend to normalise disfluencies and mispronunciations [17, 18]. This behaviour is counterproductive for detecting children’s reading errors—such as substitutions, omissions, and repetitions—which are essential diagnostic markers for literacy development [19, 20]. Recent research suggests that current models systematically underperform on such disfluent speech, effectively *erasing* the very behaviours teachers need to monitor

[21].

Furthermore, children’s speech introduces distinct acoustic variability, including unstable prosody and inconsistent articulation [22, 23]. These difficulties are compounded in low-resource contexts like Basque, where complex morphology and a lack of large-scale child-specific corpora limit model generalisation [24, 25]. While recent studies show that lightweight adaptation methods like LoRA can improve performance in specific languages [26], achieving literal transcription in a real-time streaming classroom environment remains an open technical hurdle [27, 28].

While prior work has examined ASR adaptation for low-resource languages, the faithful transcription of children’s disfluent speech remains underexplored. This work addresses this gap by adapting an autoregressive ASR model to produce unnormalized Basque transcriptions using a multi-stage fine-tuning approach with real and synthetic data.

3. System Overview

The proposed system, depicted in Figure 1, is an educational application designed to support children’s reading development in Basque through the automatic analysis of reading-aloud exercises. The application is implemented in Unity, enabling deployment on multiple platforms, including standard devices and a virtual reality environment, which is essential to maintain the engagement of the target demographic (children ages 6 to 10) and to provide vocalised audio feedback through the Orai text-to-speech (TTS) engine. The core processing logic, including the Automatic Speech Recognition (ASR) and error detection modules, is hosted in a Dockerized backend to ensure scalability and robust performance.

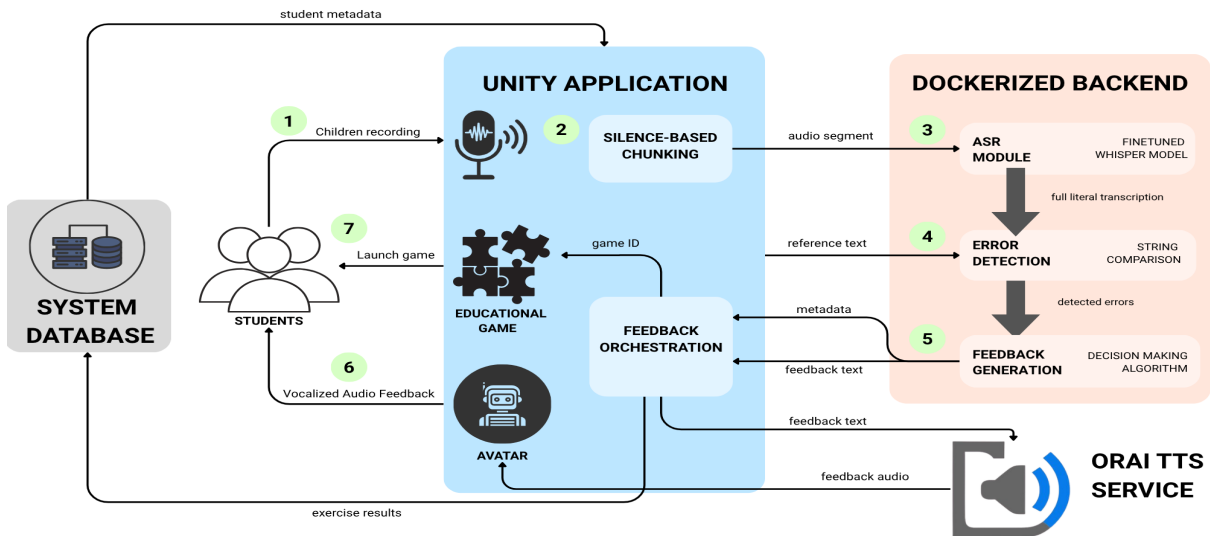


Figure 1: Conceptual architecture of the automated reading tutor system

The system consists of several interconnected modules. An automatic speech recognition (ASR) component is used to generate literal transcriptions of the children’s oral reading in Basque. This step is particularly critical, as the quality and fidelity of the transcription directly affect all subsequent processing stages. An error detection module compares the generated transcription with the reference text that the child is expected to read. Through this comparison, the system identifies different types of reading errors, such as substitutions, omissions or insertions. A decision-making module analyses the detected errors in order to determine appropriate pedagogical actions. These may include generating feedback, highlighting specific difficulties, or adapting future reading exercises to the needs of each learner. Beyond assessment and feedback, the application incorporates adaptive educational games that provide targeted practice based on each learner’s identified reading difficulties. In a classroom scenario, students interact with the system by reading short texts aloud, receiving immediate feedback through an avatar, while teachers can monitor detected reading errors and adapt instruction accordingly.

The interaction begins when a student logs in (see Figure 2 left), selects a text, and starts reading aloud (see Figure 2 middle). During the reading process, the audio signal is continuously recorded in Unity and segmented into chunks based on silence detection, following a streaming-orientated approach to minimise latency [27]. These audio segments, along with relevant metadata from the system database, are then sequentially sent to the Dockerized ASR module, where each chunk is transcribed independently.

The ASR module is responsible for generating literal transcriptions of the child's speech. Unlike standard commercial ASR solutions, which typically apply normative filtering to produce fluent text, our module is specifically optimised for literal transcription to ensure that phonetic deviations remain visible for pedagogical assessment. This requirement is critical, as the preservation of reading errors directly impacts subsequent analysis. Once the reading exercise is completed, the error detection module performs an alignment between the fully produced transcription and the reference text to identify insertions, omissions, and substitutions. This interaction loop is designed to provide immediate, engaging feedback to the child while keeping the teacher in the decision-making process.

The error detection module is implemented as a sequence alignment process between the ASR hypothesis and the reference text. This alignment is computed using Python's `difflib` library, which is based on a variation of the Ratcliff–Obershelp algorithm [29] for longest common subsequence matching. The method operates at the token level and produces a sequence of edit operations that transform the reference text into the recognised transcription.

From this alignment, four types of reading errors are identified: substitutions (incorrect word in place of the target), insertions (additional words not present in the reference), omissions (missing words), and rotations (local reordering events where words are produced in an incorrect sequence). This approach enables a direct and interpretable mapping between the produced utterance and the expected text, providing a structured representation of reading behaviour suitable for downstream pedagogical analysis.

All detected errors and reading metrics are stored in the database and processed by a decision-making module that maps detected error types (substitutions, omissions, etc.) to specific pedagogical strategies using a rule-based expert system developed in collaboration with PT specialists. This module generates personalised feedback in the form of a text string that will be transmitted back to Unity along with the list of errors and a recommended game ID. Upon reception, the application utilises the *OraiNLP Basque TTS system* to vocalise the feedback through an animated avatar, providing the student with immediate, engaging guidance.

Finally, the system recommends targeted learning activities in the form of educational games (right image in Figure 2). The pedagogical logic behind these recommendations and the feedback generation has been rigorously researched and developed in collaboration with education professionals, including classroom teachers and Therapeutic Pedagogy (PT) specialists. These activities are designed to implicitly address the most frequent errors made during the exercise, thereby reinforcing specific reading skills in an adaptive manner based on expert-defined criteria.

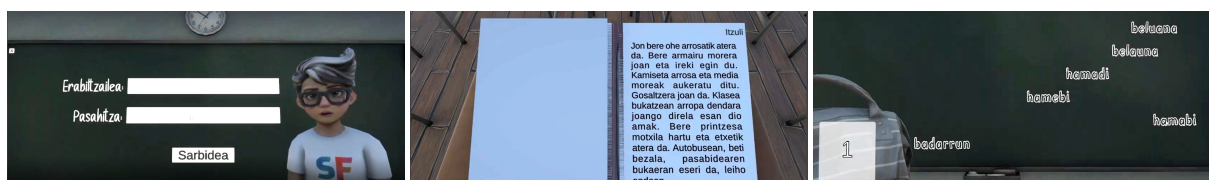


Figure 2: User interface: login screen with the avatar (left), reading exercise (middle) and an educational game (right).

4. Data

The system relies on three complementary types of speech data to support acoustic modelling, error modelling, and real-world evaluation. All audio is stored in .wav format at 16 kHz and contains only Basque speech. Transcriptions are normalised to lowercase and without punctuation to ensure consistency across datasets. These datasets (see Table 1) are selected to jointly address acoustic variability, literal transcription, and domain mismatch between adult speech and children’s reading scenarios.

Table 1

Summary of the datasets used for ASR training and evaluation.

Corpus	Core role	Transcription	Task
Adult	Acoustic modelling	Normalised	Fine-tuning, Evaluation
Synthetic	Controlled error modelling	Literal	Fine-tuning, Evaluation
Children	Real-world evaluation	Literal	Evaluation

Acoustic Modelling (Adult Speech). A large-scale composite corpus of Basque adult speech is used as a baseline, publicly available on the Hugging Face platform.¹ It combines several publicly available datasets, including the Basque subset of Mozilla Common Voice, recordings from the Basque Parliament, and curated read speech resources. This corpus primarily consists of fluent, spontaneous speech with non-literal transcriptions in which disfluencies are typically normalised. Although it does not directly correspond to the target task, the dataset provides a strong linguistic foundation that enables the model to learn general phonetic and lexical characteristics of Basque under relatively stable acoustic conditions. In this study, the corpus is mainly used for representation learning, allowing the model to acquire robust phonetic and lexical representations prior to adaptation to more specialized and challenging domains.

Controlled Error Modelling (Synthetic Data). To mitigate the limited availability of children’s reading-aloud data, a synthetic corpus is generated using the TTS system described in Section 3. The resulting dataset comprises several hundred hours of speech synthesized from a large and diverse set of sentences and multiple artificial speakers, ensuring variability in both lexical content and speaker characteristics. The utterances are designed to resemble a reading-style delivery and incorporate controlled, artificially introduced errors, including substitutions, omissions, and repetitions, to simulate common reading inaccuracies.

Errors are systematically introduced at both the word and sentence levels following predefined transformation rules that simulate typical reading difficulties. These include phonological substitutions, repetition at error points, and sentence concatenation effects. This controlled setup helps expose the model to typical reading errors under well-defined conditions.

Real-World Evaluation (Children’s Speech). A small corpus of real children’s reading speech is collected in collaboration with a Basque-medium primary school. The recordings consist of children reading short pedagogical texts, and the transcriptions are manually annotated at a literal level, preserving reading errors and disfluencies. The texts are specifically designed for early readers and reflect pedagogical progression in reading difficulty. All transcriptions are manually verified by experts, ensuring high annotation reliability and precise alignment with the spoken signal. Despite its limited size, this dataset represents the target application scenario and serves as the main benchmark for evaluating the system’s performance on real child speech.

These datasets define three complementary learning regimes: (i) acoustic modelling from adult speech, (ii) controlled error modelling through synthetic data, and (iii) real-world evaluation on children’s

¹https://huggingface.co/datasets/asierhv/composite_corpus_eu_v2.1

reading speech. The adult corpus serves as an '*acoustic baseline*' for the language and reflects general ASR performance in Basque. The synthetic corpus acts as a '*behavioural baseline*' for error detection, enabling a controlled analysis of literal transcription. Finally, the children's corpus captures the acoustic and linguistic variability of real-world reading and is used both for domain adaptation and as the primary evaluation benchmark. Together, they support the progressive adaptation of the ASR module to the application requirements. This corpus will be used for evaluation purposes due to its limited size. Table 2 provides an overview of the volume and the specific roles of each corpus within the training and evaluation framework.

Table 2

Extended summary of the corpora used for ASR training and evaluation.

Corpus	Source	Hours	Primary Characteristics
Adult Composite	Common Voice, Parliament, SLR	445.0	Spontaneous, Normalised Transcriptions
Synthetic (TTS)	OraiNLP	317.0	Synthetic Child-like, Error-infused Literal Transcriptions
Ikastola (SF)	San Fermin Ikastola	5.0	Real Child Reading, Expert-verified Literal Transcriptions

5. Model Adaptation for Educational Use

The ASR model is adapted to meet the requirements of the educational application, namely literal transcription and robustness to children's speech. Rather than focusing on benchmarking multiple architectures, the goal is to analyse how a strong pre-trained model can be adapted to these constraints through targeted fine-tuning strategies.

Autoregressive model (Whisper-small)². Whisper-small is a pre-trained encoder-decoder ASR model that has shown strong multilingual performance. Its multilingual training and large-scale weakly supervised pre-training make it particularly suitable for low-resource scenarios such as Basque. Its autoregressive decoding introduces a tendency toward linguistically well-formed outputs, which allows the model to incorporate implicit linguistic knowledge, resulting in stable and coherent transcriptions. However, this same property can introduce a bias toward normalised or linguistically well-formed outputs, potentially reducing sensitivity to disfluencies and reading errors, which are critical in the context of this work. To mitigate this effect, the training process incorporates synthetic data containing controlled disfluencies and reading errors, encouraging the model to better align with non-canonical speech patterns and preserve fidelity to the input signal.

5.1. Fine-tuning strategy

The model was adapted to the target task through a two-stage fine-tuning process. The goal of this strategy is to shift the model toward acoustic fidelity rather than linguistic smoothing.

First, Whisper-small was fine-tuned on a large corpus of adult speech, allowing it to learn robust acoustic and linguistic representations of Basque under relatively stable conditions. In a second stage, the model was further fine-tuned using the synthetic speech corpus generated through the TTS system. This dataset includes controlled reading-style utterances with artificially introduced errors such as substitutions, omissions, and repetitions. The inclusion of this data aims to expose the model to non-canonical speech patterns that are representative of children's reading behaviour. The children's speech corpus was not used during training and was reserved exclusively for evaluation. This setup allows for a clear assessment of the model's ability to generalise to real child speech without direct exposure during fine-tuning.

²<https://huggingface.co/openai/whisper-small>

5.2. Evaluation framework

To better understand the system behaviour, we analyse it under three complementary scenarios: acoustic modelling (adult speech), controlled error modelling (synthetic data), and real-world evaluation (children’s speech). Each condition provides insight into a different aspect of model behaviour. Adult speech reflects the ability to model Basque linguistic structure, synthetic speech allows controlled assessment of transcription behaviour under known error patterns, and children’s speech captures robustness to real-world acoustic variability.

Incorporating synthetic data allows the model to better capture non-canonical speech patterns, aligning its behaviour with the application’s requirements. The model fine-tuned on both adult and synthetic speech is better suited than the version trained only on adult speech. In particular, the behaviour on children’s speech improves despite the absence of child speech during training, indicating that the synthetic data effectively bridges part of the domain gap.

These findings suggest that introducing controlled disfluencies during training helps the model better handle irregular speech patterns, improving its ability to generalise to the target task while maintaining transcription fidelity.

6. Discussion

The observations obtained during system development highlight a fundamental mismatch between the objectives of standard ASR systems and the requirements of pedagogical applications. While the autoregressive Whisper-small model provides acoustically robust and linguistically coherent transcriptions, its decoding strategy introduces a systematic normalisation bias that tends to smooth reading irregularities. This behaviour is detrimental in educational contexts, where disfluencies such as substitutions, omissions, and repetitions constitute the primary diagnostic signals for assessing reading development.

The proposed multi-stage fine-tuning strategy, incorporating error-infused synthetic data, partially mitigates this limitation by encouraging the model to better preserve non-canonical speech patterns. Notably, the observed improvements on real children’s speech, despite the absence of such data during training, suggest that controlled synthetic errors can effectively bridge part of the domain gap. This finding supports the use of synthetic data not only as a data augmentation strategy, but as a mechanism to explicitly shape model behaviour toward task-specific objectives.

Beyond the ASR component, these observations have direct implications for the overall MentorIA system. The application relies on a tightly coupled pipeline in which transcription quality directly conditions the performance of downstream modules, particularly error detection and pedagogical decision-making. In this context, transcription errors are not isolated artefacts but propagate through the system, potentially leading to incorrect identification of reading difficulties and suboptimal feedback. Therefore, optimising ASR for literal fidelity is not only a modelling challenge but a system-level requirement for reliable educational support.

However, several limitations remain. The size of the real children’s speech corpus is limited, constraining the evaluation of generalisation in highly variable acoustic conditions. The use of synthetic data relies on predefined error distributions, which may not fully capture the complexity of natural reading behaviour. Finally, the current alignment-based error detection approach may be sensitive to transcription noise, highlighting the need for more robust methods.

7. Future Work

Future development of the MentorIA project will extend along several complementary research directions, targeting improvements at the data, modelling, error detection, and analysis levels. From a data perspective, future work will explore more advanced data augmentation strategies to better capture the acoustic variability of children’s speech. In addition to pitch shifting and vocal tract length normalisation (VTLN), we plan to incorporate perturbations such as tempo variation, noise injection,

and prosodic manipulation. Furthermore, the integration of richer acoustic representations will be investigated, including alternative feature spaces such as log-Mel variants or self-supervised speech embeddings, with the goal of improving robustness in low-resource conditions.

With regard to the modelling approach, while the current system relies on a fine-tuned autoregressive Whisper architecture, alternative ASR paradigms that may better preserve disfluencies will be explored. In particular, non-autoregressive and hybrid architectures, such as CTC-based or transducer models, will be evaluated to mitigate the normalisation bias observed in autoregressive systems. Additionally, lightweight adaptation techniques and modular architectures will be considered to improve domain adaptation efficiency and facilitate deployment in real-time educational environments.

For error detection, current alignment-based methods will be extended toward more robust and flexible approaches. Future work will investigate learning-based alignment models and error classification systems capable of handling highly disfluent input and partial mismatches between hypothesis and reference. This includes the development of models that jointly leverage acoustic and textual information, enabling more accurate detection of subtle reading behaviours and reducing sensitivity to ASR errors.

Finally, the system will evolve beyond raw text-based analysis by incorporating richer speech-derived features. In particular, future extensions will focus on modelling prosody, rhythm, and intonation patterns, as well as non-lexical phenomena such as pauses, hesitations, and blocks. These features are essential for assessing reading fluency and can provide complementary diagnostic information that is not captured by text alone, enabling a more comprehensive characterisation of reading performance. Overall, MentorIA represents an ongoing effort to bridge speech technology and educational practice, with a focus on real-world deployment in Basque classrooms.

Acknowledgments

This research is part of the project “*MentorIA: Desarrollo cognitivo infantil tutorizado*” (Ref: 011-1365-2024-000121). The project is funded by the Government of Navarre through the Department of Economic and Business Development (General Directorate of Industry, Energy, and S3 Strategic Projects) and co-financed by the European Union through the European Regional Development Fund (ERDF). The consortium includes the collaborating entities *Sistemas Navarra S.A.* and *San Fermín Ikastola S. Coop. Ltda.*, with the participation of the *Universidad Pública de Navarra (UPNA)*, specifically the *Distributed Systems Group*.

Declaration on Generative AI

During the preparation of this work, the authors used Claude (Anthropic) to assist with LaTeX formatting, including the preparation and formatting of tables, figure placement, captions, and related document elements. ChatGPT (OpenAI) was used to refine the English writing, improve grammar and spelling, and enhance the clarity, fluency, and readability of the manuscript. Following the use of these tools, the authors carefully reviewed and revised the generated suggestions and take full responsibility for the final content of the publication.

References

- [1] R. Prabhavalkar, T. Hori, T. N. Sainath, R. Schlüter, S. Watanabe, End-to-end speech recognition: A survey, *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 32 (2023) 325–351.
- [2] M. Malik, M. K. Malik, K. Mehmood, I. Makhdoom, Automatic speech recognition: a survey, *Multimedia Tools and Applications* 80 (2021) 9411–9457.
- [3] S. Singh, M. Singh, V. Kadyan, Speech recognition transformers: topological-lingualism perspective, *arXiv preprint arXiv:2408.14991* (2024).
- [4] S. Latif, A. Zaidi, H. Cuayahuitl, F. Shamshad, M. Shoukat, J. Qadir, Transformers in speech processing: A survey, *arXiv preprint arXiv:2303.11607* (2023).

- [5] A. Baevski, Y. Zhou, A. Mohamed, M. Auli, wav2vec 2.0: A framework for self-supervised learning of speech representations, *Advances in neural information processing systems* 33 (2020) 12449–12460.
- [6] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, Robust speech recognition via large-scale weak supervision, in: *International conference on machine learning*, PMLR, 2023, pp. 28492–28518.
- [7] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. Von Platen, Y. Saraf, J. Pino, et al., Xls-r: Self-supervised cross-lingual speech representation learning at scale, *arXiv preprint arXiv:2111.09296* (2021).
- [8] H. Yadav, S. Sitaram, A survey of multilingual models for automatic speech recognition, *arXiv preprint arXiv:2202.12576* (2022).
- [9] H. Kheddar, M. Hemis, Y. Himeur, Automatic speech recognition using advanced deep learning approaches: A survey, *Information fusion* 109 (2024) 102422.
- [10] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, et al., Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups, *IEEE Signal Processing Magazine* 29 (2012) 82–97.
- [11] L. R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, *Proceedings of the IEEE* 77 (1989) 257–286. doi:10.1109/5.18626, errata at <http://alumni.media.mit.edu/rahimi/rabiner/rabiner-errata/rabiner-errata.html>.
- [12] G. E. Dahl, D. Yu, L. Deng, A. Acero, Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition, *IEEE Transactions on Audio, Speech, and Language Processing* 20 (2012) 30–42. doi:10.1109/TASL.2011.2134090.
- [13] G. Skantze, A. Hjalmarsson, Towards incremental speech generation in conversational systems, *Computer Speech & Language* 27 (2013) 243–262. URL: <https://www.sciencedirect.com/science/article/pii/S0885230812000411>. doi:<https://doi.org/10.1016/j.csl.2012.05.004>, special issue on Paralinguistics in Naturalistic Speech and Language.
- [14] V. Srivastav, S. Zheng, E. Bezzam, E. L. Bihan, N. R. Koluguri, P. Želasko, S. Majumdar, A. Moumen, S. Gandhi, Open asr leaderboard: Towards reproducible and transparent multilingual and long-form speech recognition evaluation, 2026. URL: <https://arxiv.org/abs/2510.06961>. arXiv:2510.06961.
- [15] M. Bain, J. Huh, T. Han, A. Zisserman, WhisperX: Time-Accurate Speech Transcription of Long-Form Audio, in: *Proc. INTERSPEECH 2023*, 2023, pp. 5556–5560. doi:10.21437/Interspeech.2023-2287.
- [16] K. Kuhn, V. Kersken, B. Reuter, N. Egger, G. Zimmermann, Measuring the accuracy of automatic speech recognition solutions, *ACM Transactions on Accessible Computing* 16 (2024) 1–23.
- [17] P. J. Lou, M. Johnson, End-to-end speech recognition and disfluency removal, *arXiv preprint arXiv:2009.10298* (2020).
- [18] D. Mujtaba, N. R. Mahapatra, M. Arney, J. S. Yaruss, H. Gerlach-Houck, C. Herring, J. Bin, Lost in transcription: Identifying and quantifying the accuracy biases of automatic speech recognition systems against disfluent speech, *arXiv preprint arXiv:2405.06150* (2024).
- [19] E. Goikoetxea, Reading errors in first-and second-grade readers of a shallow orthography: Evidence from spanish, *British Journal of Educational Psychology* 76 (2006) 333–350.
- [20] M. De Rom, M. Van Reybroeck, Guessing errors made by children with dyslexia in word and text reading, *Frontiers in Psychology* 15 (2024) 1195696.
- [21] P. Mihajlik, Y. Meng, M. S. Kadar, J. Linke, B. Schuppler, K. Mády, On disfluency and non-lexical sound labeling for end-to-end automatic speech recognition, *Interspeech 2024* (2024) 1270–1274.
- [22] A. Potamianos, S. S. Narayanan, S. Lee, Automatic speech recognition for children., in: *Eurospeech*, 1997.
- [23] H. K. Kathania, S. R. Kadiri, P. Alku, M. Kurimo, Study of formant modification for children asr, in: *ICASSP 2020, IEEE*, 2020, pp. 7429–7433.
- [24] I. Odriozola, et al., Basque speecon-like and basque speechdat mdb-600: speech databases for the development of asr technology for basque., in: *LREC*, 2014, pp. 2658–2665.
- [25] O. Kjartansson, et al., Open-source high quality speech datasets for basque, catalan and galician,

- in: Proceedings of SLTU and CCURL, 2020, pp. 21–27.
- [26] H. Polat, et al., Implementation of a whisper architecture-based turkish asr system and evaluation of fine-tuning with lora adapter, *Electronics* 13 (2024) 4227.
 - [27] H. Miao, et al., Transformer-based online ctc/attention end-to-end speech recognition architecture, in: *ICASSP 2020, IEEE*, 2020, pp. 6084–6088.
 - [28] R. Southwell, et al., Challenges and feasibility of automatic speech recognition for modeling student collaborative discourse in classrooms, *International Educational Data Mining Society* (2022).
 - [29] J. W. Ratcliff, D. Metzener, Pattern matching: The gestalt approach, *Dr. Dobb's Journal* 46 (1988).