3rd Edition

Data Research meetup by MagIC

WAVe: Word-Aligned Verification of Synthetic Speech for Automatic Speech Recognition

Yuriy Perezhohin (yperezhohin@novaims.unl.pt)

NOVA Information Management School (NOVA IMS), Universidade Nova de Lisboa, Campus de Campolide, 1070-312 Lisboa, Portugal;

REMYND, Alameda Bonifácio Lázaro Lozano nº15, 1ºC, 2780-125, Oeiras, Portugal

INTRODUCTION

Automatic Speech Recognition (ASR) for low-resource languages often relies on synthetic speech generated by pairing Large Language Model (LLM) transcripts with Text-to-Speech (TTS) generated audio. However, blindly incorporating synthetic data can introduce mispronunciations, word omissions, and prosodic anomalies that degrade ASR performance and increase training time [1], [2]. Previous filtering methods assessed audio-text similarity at the sentence level, which masks localized word-level errors, a synthetic utterance may appear semantically aligned while still containing critical defects [3].

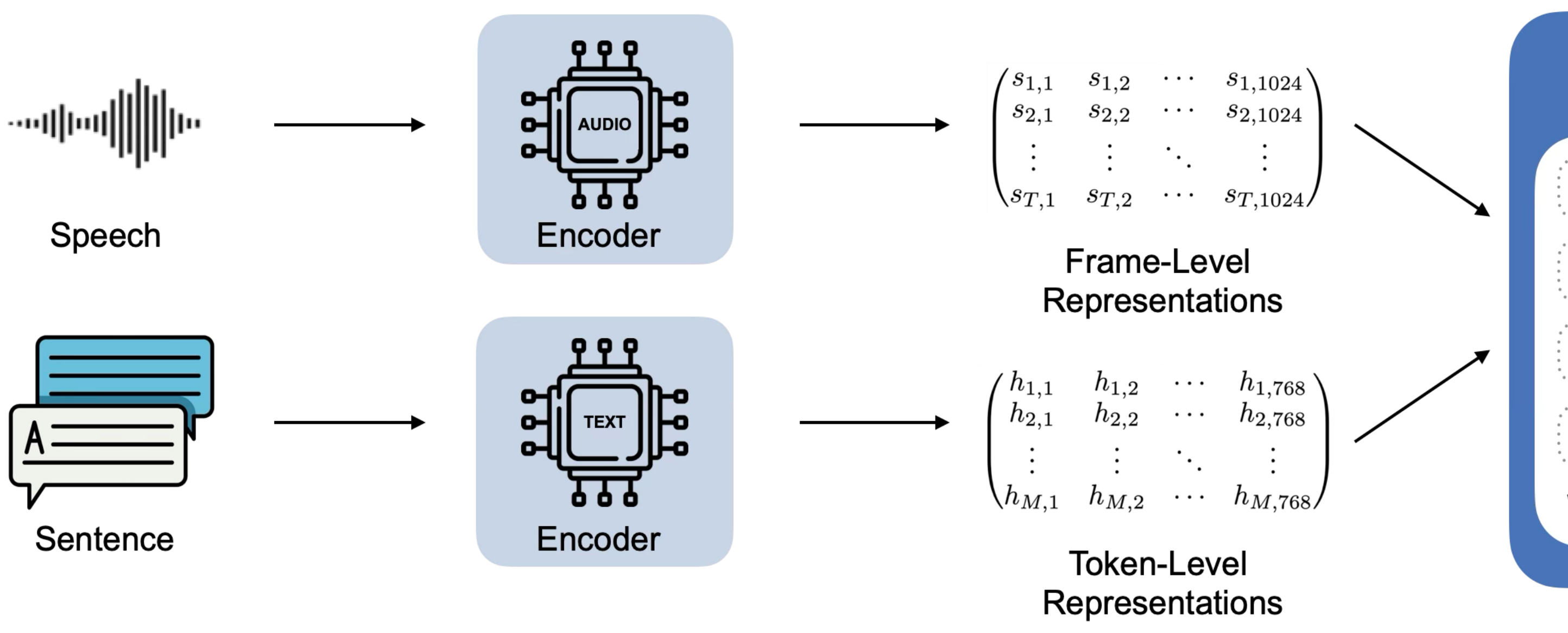
We present **WAVe** (Word-Aligned Verification), a multimodal embedding model that verifies each word against its corresponding audio frame through attention-based alignment. A Gated Linear Unit (GLU) scorer assigns confidence scores to each word, enabling fine-grained quality assessment.

METHODS AND MATERIALS

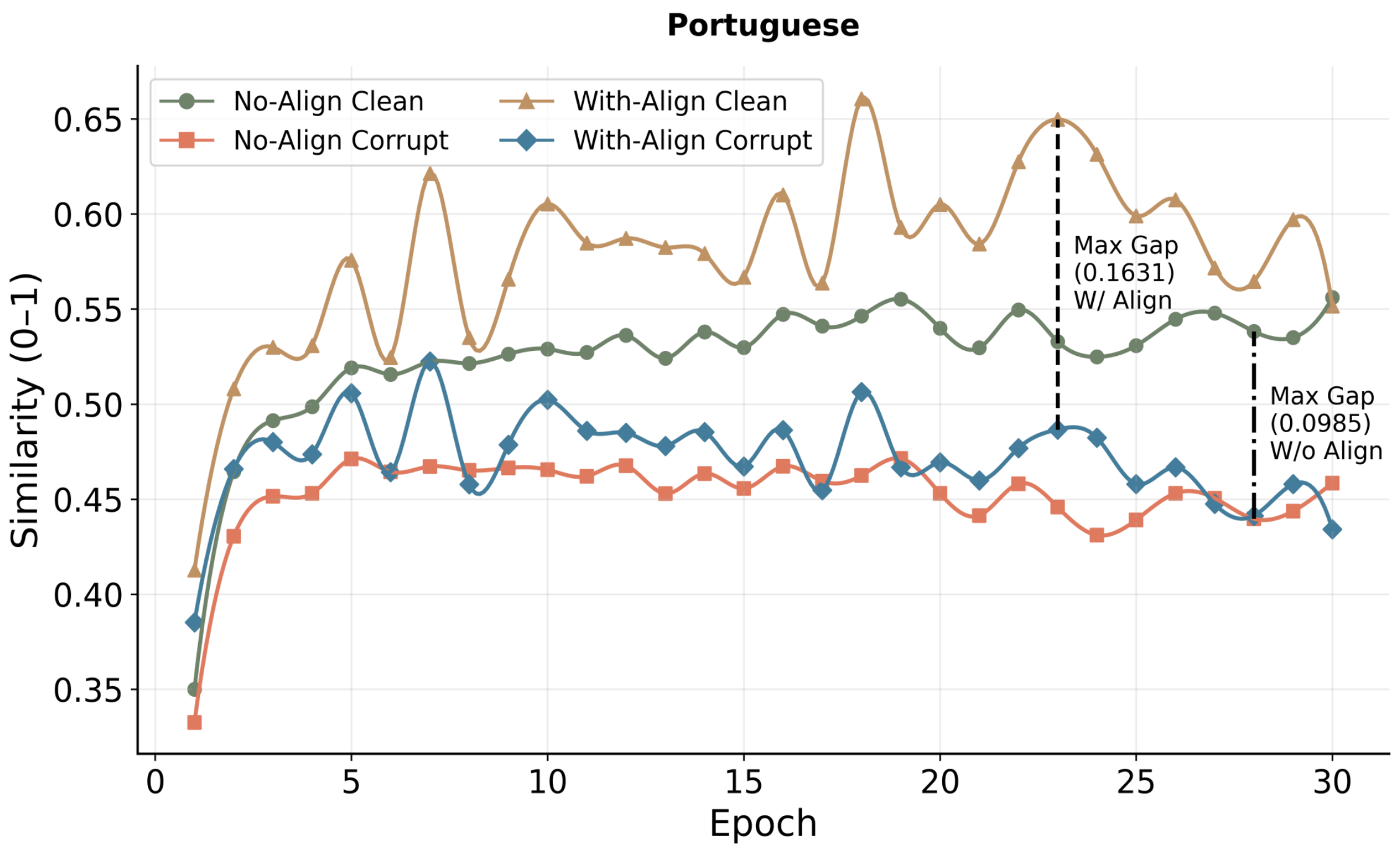
Our methodology follows a four-stage pipeline:

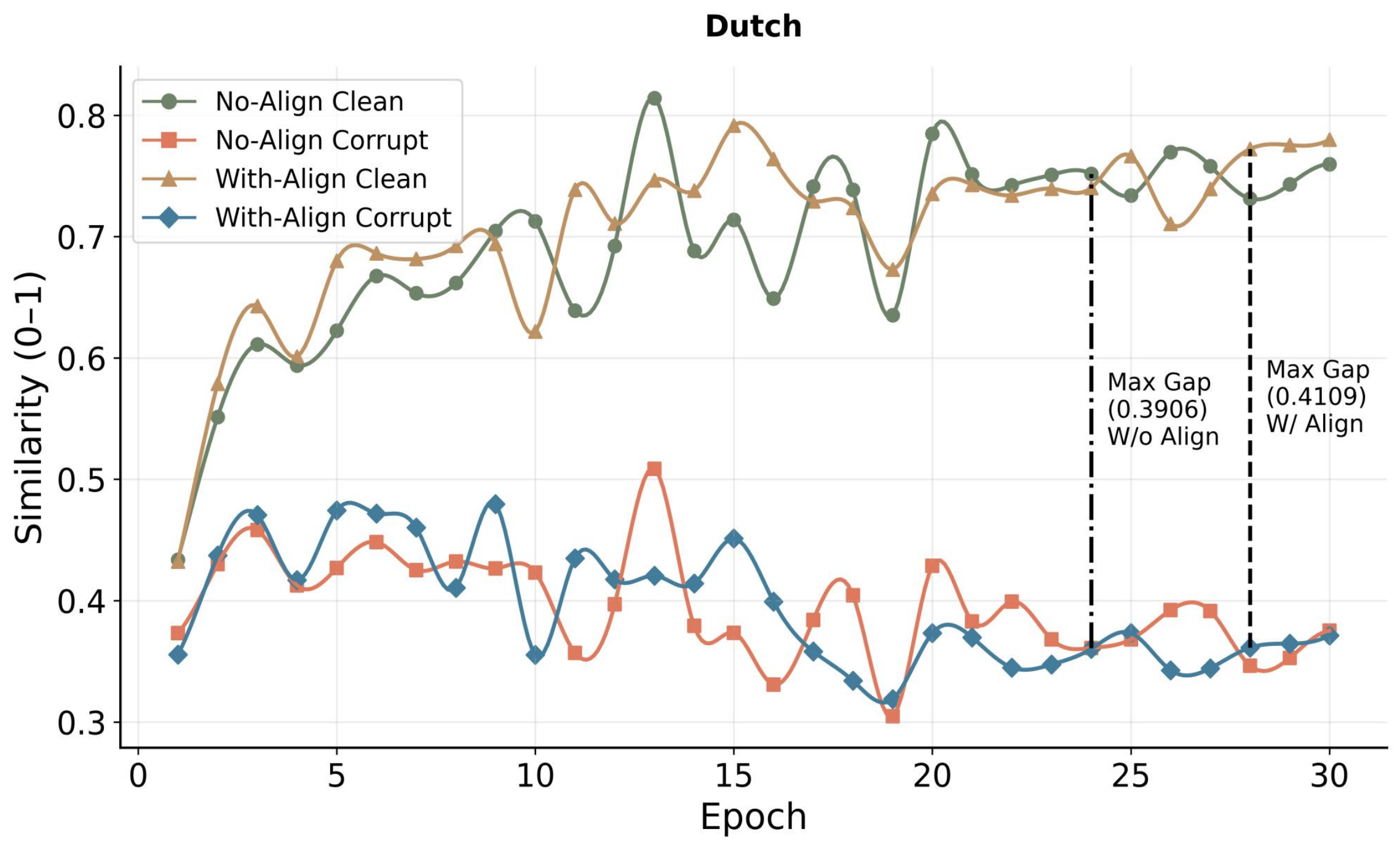
1. *WAVe Training* — We train WAVe via contrastive learning on Common Voice, contrasting clean audio-text pairs with corrupted versions to learn alignment.
2. *Synthetic Generation* — We then generate synthetic data using GPT-4o-mini transcripts and OpenAI's TTS (22k Portuguese, 35k Dutch samples).
3. *Quality Filtering* — The trained WAVe scores each pair (q ∈ [0,1]), partitioning them into subsets as shown in the table above.
4. *ASR Evaluation* — Finally, we fine-tune Whisper (Tiny, Small, Large-v3) on filtered data and evaluate on Common Voice and MLS benchmarks.

The WAVe architecture (right) illustrates the flow from encoders through cross-modal attention to word-level confidence scoring.


Speech
AUDIO
Encoder
Sentence
TEXT
Encoder
$\begin{pmatrix} s_{1,1} & s_{1,2} & \cdots & s_{1,1024} \\ s_{2,1} & s_{2,2} & \cdots & s_{2,1024} \\ \vdots & \vdots & \ddots & \vdots \\ s_{T,1} & s_{T,2} & \cdots & s_{T,1024} \end{pmatrix}$
Frame-Level Representations
$\begin{pmatrix} h_{1,1} & h_{1,2} & \cdots & h_{1,768} \\ h_{2,1} & h_{2,2} & \cdots & h_{2,768} \\ \vdots & \vdots & \ddots & \vdots \\ h_{M,1} & h_{M,2} & \cdots & h_{M,768} \end{pmatrix}$
Token-Level Representations
Word to Frame Attention
GLU Scorer
Given
this
good
model
0.90
0.50
0.83
0.91
Word - N
Frame - N
Score - N

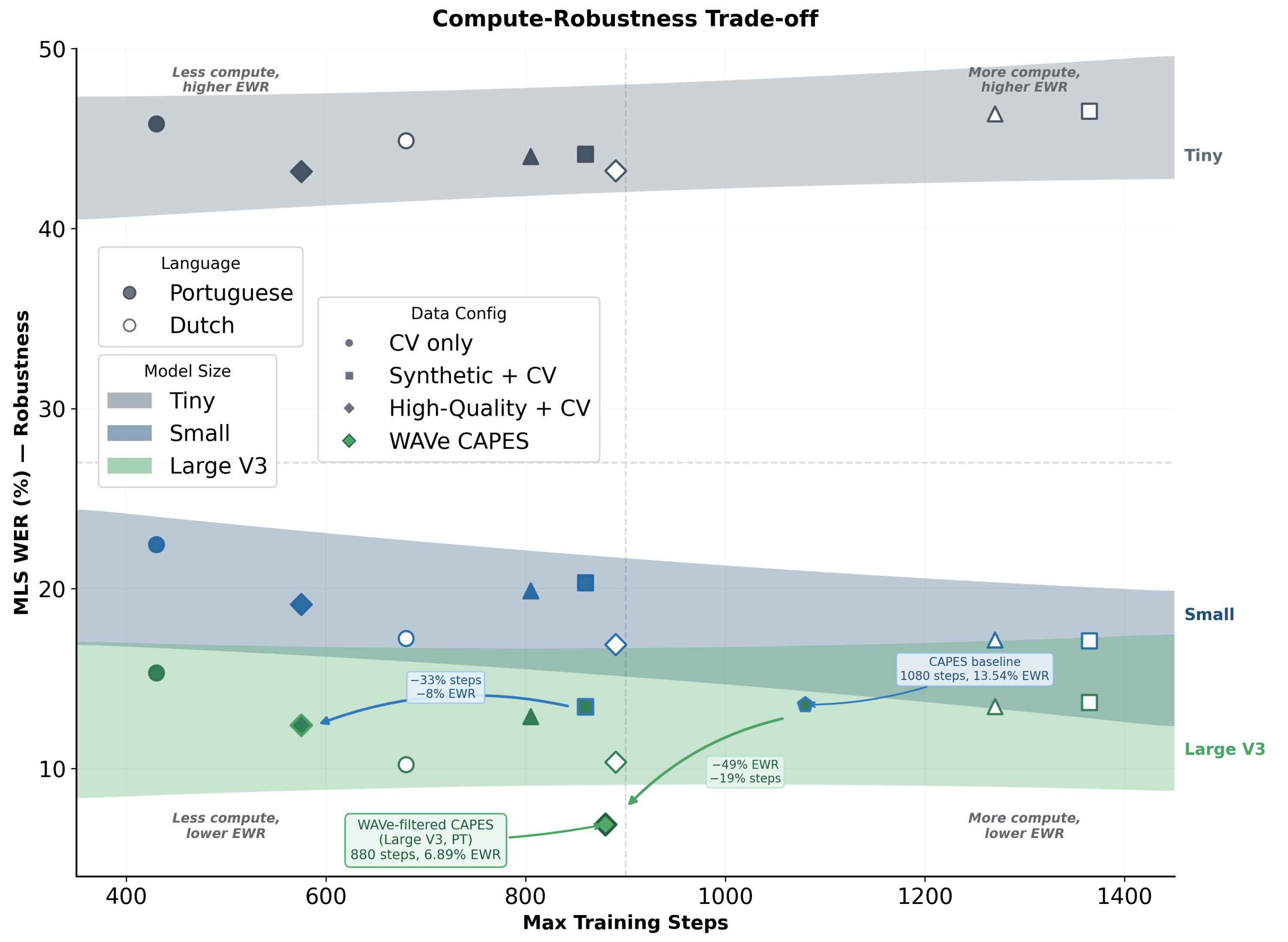
RESULTS & DISCUSSION





WAVe Training Performance

Word-level alignment supervision notably improves the model's ability to distinguish clean from corrupted pairs. For Portuguese, figure on the left, the clean-corrupt similarity gap increases from 9.85% (without alignment) to 16.31% (with alignment), a 6.5% absolute improvement achieved 5 epochs earlier. The Dutch language, figure on the right shows larger absolute margins (41.09%gap) due to its 60% larger training corpus.



ASR Performance

WAVe demonstrates strong performance for Whisper Large-V3 on Portuguese across both evaluation benchmarks. On Common Voice (in-domain), high-quality filtering achieves 7.94% WER compared to 8.33% with unfiltered data, while requiring only 575 training steps, a 33% reduction. This efficiency stems from removing low-quality samples that introduce conflicting acoustic signals, allowing faster convergence. The benefits become even more pronounced for cross-domain generalization on MLS. WAVe-filtered CAPES dataset reduces WER from 13.54% to 6.89%, using 19% fewer steps and 30% less data. This substantial gain indicates that word-level filtering retains samples with robust acoustic representations that transfer well across domains, rather than overfitting to synthesis artifacts and hindering performance.

CONCLUSION

Our approach achieves a **34% average reduction** in training steps while simultaneously decreasing WER, demonstrating that quality-based filtering delivers both computational efficiency and improved accuracy. Most critically, WAVe exhibits strong robustness: cross-domain evaluation reveals up to 49% improvement in generalization (13.54% → 6.89% MLS WER) using 30% less data. By retaining only high-quality synthetic samples, our method ensures models learn transferable acoustic representations, establishing quality-over-quantity as the key paradigm for efficient low-resource ASR augmentation.

REFERENCES

- [1] Dhamyal et al. (2024). Using Voicebox-based Synthetic Speech for ASR Adaptation. *Proc. SynData4GenAI*.
- [2] Wang et al. (2025). From Tens of Hours to Tens of Thousands: Scaling Back-Translation for Speech Recognition. *arXiv*.
- [3] Manco et al. (2022). Learning music audio representations via weak language supervision. ICASSP 2022.

ACKNOWLEDGEMENTS

This work was supported by national funds through FCT (Fundação para a Ciência e a Tecnologia), under the project - UID/04152/2025 - Centro de Investigação em Gestão de Informação (MagIC)/NOVA IMS -https://doi.org/10.54499/UID/04152/2025 (2025-01-01/2028-12-31) and UID/PRR/04152/2025 https://doi.org/10.54499/UID/PRR/04152/2025 (2025-01-01/ 2026-06-30). This work was funded by the European Union throught the project 101084013 - DIGITAL4Business. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HADEA). Neither the European Union nor the granting authority can be held responsible for them. A special thank you to the Professor Mauro Castelli for his support to develop this work

Funded by:

NOVA IMS MagIC INFORMATION MANAGEMENT RESEARCH CENTER
NOVA IMS INFORMATION MANAGEMENT SCHOOL
fct Fundação para a Ciência e a Tecnologia
PRR Plano de Recuperação e Resiliência
REPÚBLICA PORTUGUESA
Financiado pela União Europeia NextGenerationEU

NOVA Information Management School, 18 December 2025

ONLINE VERSION