A Comparative Study of Speech-to-Text for Italian*

Michela Quadrini[1], Michele Loreti[1], Mattia Luciani[1,2], Ivano Corradetti[2], Matteo Carboni[2] and Stefano Vagnoni[2,*,†]

[1]*School of Science and Technology, University of Camerino, Via Madonna delle Carceri, 9, Camerino, 62032, Italy*
[2]*INTELLIGENZA AUMENTATA SRL, via dell'Aspo 42, Ascoli Piceno*

Abstract

Accurate and efficient speech recognition devices is essential for enabling real-time applications in various fields, including healthcare. One example is the support for individuals with neurodegenerative conditions such as dementia, where automatic speech recognition can assist in monitoring cognitive decline and facilitating remote assessments. Traditional diagnostic methods often rely on in-person questionnaires, which can be time-consuming and demanding for both patients and caregivers.

In this study, we present a comparative analysis aimed at identifying the most effective combination of speech-to-text systems and spelling correction methods suitable for real-time applications on devices with limited hardware capabilities. We evaluate three speech-to-text systems, Whisper Tiny, Whisper Base, and Vosk,alongside two spelling correction approaches. All models were tested using the publicly available Italian Common Voice dataset. Results indicate that Whisper Base consistently outperforms both Whisper Tiny and Vosk. Additionally, in low-noise environments, the impact of spelling correction on Whisper's performance is minimal, suggesting its robustness even without post-processing.

Keywords

Speech-to-text Systems Evaluation, spelling correction, Dementia.

1. Introduction

Accurate and efficient speech recognition is becoming increasingly important for enabling real-time applications in a variety of fields, including education, accessibility, and healthcare. In the medical domain, automatic speech recognition (ASR) technologies can play a key role in remote assessments, longitudinal monitoring, and reducing the burden of traditional diagnostic procedures. A relevant example is the support for individuals with neurodegenerative conditions such as dementia, where speech-based tools can assist in monitoring cognitive decline and enabling non-invasive screening solutions.

Dementia is a syndrome characterized by a progressive decline in cognitive function that impairs daily functioning. Currently, over 55 million people worldwide live with dementia, with nearly 10 million new cases diagnosed each year, and more than 60% of those affected residing in low- and middle-income countries [1]. The Diagnostic and Statistical Manual of Mental Disorders (DSM-5) lists 13 possible causes of dementia [2], with Alzheimer's disease accounting for approximately 70% of cases [3]. Other causes include Lewy body disease, traumatic brain injuries, vascular issues, and frontotemporal dementia [4]. Early detection and continuous monitoring are crucial to managing symptoms and improving quality of life through both pharmacological and non-pharmacological interventions. Standard diagnostic tools, such as the Mini-Mental State Examination (MMSE), are widely used for cognitive screening and are considered the gold standard in many clinical settings. Other methods, like the AD8 [5] and Mini-Cog, are also in use. However, these tools often require administration in person, making them

Ital-IA 2025: 5th National Conference on Artificial Intelligence, organized by CINI, June 23-24, 2025, Trieste, Italy)

*Michela Quadrini

†These authors contributed equally.

michela.quadrini@unicam.it (M. Quadrini); michele.loreti@unicam.it (M. Loreti); mattia.luciani@studenti.unicam.it (M. Luciani); ivano.corradetti@cambiaverso.it (I. Corradetti); ivano.corradetti@cambiaverso.it (M. Carboni); stefano.vagnoni@gmail.com (S. Vagnoni)
0000-0003-0539-0290 (M. Quadrini); 0000-0003-3061-863X (M. Loreti)

time-consuming and potentially burdensome for patients and caregivers—especially in advanced stages or in settings with limited clinical access.

Advancements in artificial intelligence (AI) have introduced new opportunities for supporting such assessments. For instance, mobile apps like Sea Hero Quest leverage gamified navigation tasks to capture spatial orientation performance, which may decline in early dementia [6]. Wearable devices that track physical activity and sleep patterns can also offer valuable early indicators citeyang2017diagnostic. More recently, large language models (LLMs) have dramatically improved speech-to-text and text-to-speech capabilities, enabling real-time speech processing for clinical applications. These models can provide rapid, automated analyses of verbal output, helping clinicians monitor patients remotely. Nevertheless, the computational demands of these models present challenges, especially on devices with limited hardware capabilities.

This paper presents a comparative study aimed at identifying the optimal combination of speech-to-text systems and spelling correction methods for real-time applications on devices with constrained computational resources. We evaluate three widely used speech-to-text systems, i.e., Whisper-tiny, Whisper-Base [7], and Vosk API [8], along with two different spelling correction approaches. To ensure the generalizability of our findings, we train and test the systems using a publicly available Italian dataset, called Common Voice dataset. Our results demonstrate that Whisper-Base outperforms both Whisper-tiny and Vosk in terms of accuracy and efficiency. Additionally, we show that the performance of Whisper, with and without spelling correction, is comparable when background noise levels are low. Finally, we integrate Whisper into a demo application to evaluate their effectiveness in real-world scenarios, providing a proof of concept for their potential use in dementia monitoring.

The rest of the paper is organized as follows. Section 2 describes the main automatic speech recognition systems, from early-stage approaches to modern methods based on deep learning architectures. Section 3 introduces the performed experiments and the selected toolkit's main reasons. The paper ends with some conclusions and future works, Section 4.

2. Overview of Automatic Speech Recognition

Automatic speech recognition (ASR) is the technology that recognizes and converts spoken language into text. It is applied in various scenarios, such as voice assistants, media applications, and voice-command systems. ASR has witnessed remarkable progress, revolutionizing human-computer interaction across diverse domains [9]. Nevertheless, this progress has not been uniformly distributed across all languages.

In the early stages, manual feature extraction and standard approaches such as Gaussian Mixture Models and Hidden Markov Model have been developed and used. These architectures were most suitable for addressing fundamental issues, as their limitations might complicate large-scale, complex real-world problems. Recently, models such as RNNs and CNNs, as well as Transformers, were applied to ASR with considerable success. ASR models like Wav2Vec, Wav2Vec2, HuBERT, Whisper, and WavLM have significantly improved the accuracy and robustness of ASR systems, paving the way for more efficient and reliable STT transcription technologies.

Wav2Vec Wav2Vec is an end-to-end architecture that utilizes self-supervised learning, combining transformer and convolutional layers [10]. The model employs a multi-layer convolutional feature encoder to transform raw audio waveforms into latent speech representations. These latent representations are input to the network, which is transformer-masked. The self-supervised learning objective is defined by the discrete outputs generated when the transformer network quantizes the continuous representations. These quantized representations are contextualized by the transformer module's attention blocks, producing a set of discrete contextual representations. The feature encoder consists of seven convolutional blocks with 512 channels. The transformer network consists of 24 blocks, each having 1024 dimensions, 4096 inner dimensions, and a total of 16 attention heads.

DiscreteBERT DiscreteBERT is an adaptation of the BERT architecture designed to process discrete token inputs instead of continuous token embeddings [11]. This modification makes it particularly suitable for handling tokenized inputs from various sources, such as audio or text, where the tokens represent distinct information units. Unlike traditional models that operate in a unidirectional manner (either left-to-right or right-to-left), DiscreteBERT can understand the context in both directions. It uses a transformer architecture that employs self-attention mechanisms to assess the importance of each token within the sequence. Additionally, DiscreteBERT is trained using Masked Language Modeling (MLM), where the model learns to predict randomly masked tokens in the input sequence, improving its ability to capture and understand the broader context of the language.

Whisper. Whisper is a model developed by OpenAI based on integrating supervised and self-supervised learning techniques [7]. It is trained on a diverse dataset of 680,000 hours of multilingual and multitask supervised data, allowing it to learn robust representations from noisy audio. This approach performs well in transcribing speech accurately across a variety of languages and tasks, making it ideal for real-world applications. Whisper excels in challenging scenarios, improving speech recognition even in difficult conditions. It is also capable of performing multiple speech-related tasks.

WavLM WavLM is a self-supervised ASR model that learns contextual speech representations from raw audio without transcriptions [12]. It combines ASR and language modeling techniques, enabling scalable training across diverse languages and maintaining strong performance in real-world tasks

Kaldi and Vosk Kaldi is a flexible and customizable speech recognition toolkit with a rich feature set and strong community support. Vosk, built on Kaldi, is lightweight and optimized for real-time, low-latency applications on resource-constrained devices. It supports multilingual models and allows custom training, making it ideal for mobile and embedded use

2.1. Dataset

The performance of ASR systems depend on high-quality, annotated speech datasets for effective training and testing. Datasets such as the Italian Spontaneous Dialogue Dataset [13] and the Common Voice project [14] provide a wide range of authentic speech samples, allowing ASR models to manage spontaneous speech, regional accents, and background noise. These resources are valuable for advancing voice assistants, automated transcription services, and accessibility tools tailored for Italian speakers.

Common Voice Mozilla's Common Voice is an open-source dataset built through user-contributed recordings and validations [14]. As of May 2022, it includes 18.6k hours of speech data (14.1k validated) across 112 languages, with ongoing updates.

Italian Spontaneous Dialogue Dataset The Italian Spontaneous Dialogue Telephony dataset features transcribed conversations from 676 native speakers on over 20 topics, with metadata like age, gender, and region. It enhances model performance in real-world tasks and complies fully with privacy and data protection regulations.

2.2. Evaluation criteria in ASR

Different metrics have been defined to evaluate the effectiveness and suitability of ASR techniques. Specific metrics for ASR have also been established, including word error rate (WER), Word Recognition Rate (WRR), and Character error rate (CER) [15]. WER determines the ratio of incorrectly recognized words to the overall number of processed words, as follows

$$WER = \frac{S + D + I}{N} \tag{1}$$

Where I, D, S, H, and N denote the number of insertions, deletions, substitutions, hits, and input words, respectively. WRR, also known as the Word Accuracy Rate, is a variant of the WER used to measure the

Model	WER	CER	Fine-Tune	WER(Fine-Tuned)	CER(Fine-Tuned)
Vosk-Small-It	40%		No	-	-
Whisper-Tiny	62%	24%	Yes	28.3%	10.08%
Whisper-Base	45%	16%	Yes	21.7%	7.28%

Table 1
Difference between the baseline and fine-tuned models

efficiency of an ASR. The metric is computed as follows

$$WRR = \frac{N - S - D - I}{N} = \frac{H - I}{N} \tag{2}$$

$H = N - (S + D)$ is the number of correctly predicted words. The CER follows the same evaluation principle, except that it measures errors in characters rather than words. The CER value can be determined using the equation below.

$$CER = \frac{S + D + I}{N} \tag{3}$$

Where S, D, and C denote the number of substitutions, insertions, deletions, and correct characters, respectively. N is the total number of characters in the source.

3. Experiments

We conducted experiments on the Common Voice dataset, focusing exclusively on the Italian language, to identify the optimal combination of ASR systems and spelling correction methods for real-time applications on devices with constrained computational resources. We selected Whisper and Vosk models and compared their performances before and after fine-tuning, which exploits the Italian part of the Common Voice dataset. The obtained results are reported in the Table 1. In particular, we fine-tuned the Whisper Base model with 74 million parameters, the Whisper Tiny model with 39 million parameters and Vosk. Whisper utilizes a Transformer architecture for its encoder and decoder, which share an identical structural setup. In the Whisper Base and Tiny versions, the encoder and decoder consist of 16 and 6 Transformer blocks, respectively. Additionally, the convolutional layers at the encoder's front end were initialized randomly, while the encoder's CTC projection layer was initialized with weights from the word embeddings used in the decoder. During fine-tuning, all model parameters were updated using gradients. We set the batch size to 8 with a gradient accumulation of 4 to have an effective batch of 32 per step and the learning rate to 2,5e-05, which is the one advised for fine-tuning this model with a warm-up linear learning rate scheduler that included 240 warm-up steps. The fine-tuning process consists of 5 epochs, and the final evaluation uses the average of the model checkpoints from the epochs.

Vosk, suitable for working in devices with limited computational capacities, utilizes deep learning models. It combines deep neural networks, Deep Neural Network Hidden Markov Models (DNN-HMM) and RNN-LM, and Mel-frequency Cepstral Coefficients (MFCC) as an acoustic model based on Kaldi to extract features that are the input of DNN-HMM that associates MFCC to spoken language. After that, a n-gram language model combined with RNN-LM predicts the probability of a word based on the context. The fine-tuning process of a Vosk model is time-consuming. It requires updating the language model and using Kaldi scripts to recreate the acoustic model based on the dataset's audio. For this reason, Vosk is only considered as a baseline for a small and fast model, suitable for devices with limited computational power, as the results obtained from the model were acceptable.

By observing the result reported in Table 1, we observe that the Whisper-Base outperforms the others. The Vosk-Small-It model, which serves as a baseline for low-resource devices, achieved a Word Error Rate (WER) of 40%. In contrast, the Whisper-Tiny model initially showed a higher WER of 62%, indicating a relatively poor performance for general ASR tasks. However, after finetuning, the

Whisper-Tiny model significantly reduced WER, improving to 28.3%. Similarly, the Character Error Rate (CER) also improved to 10.08%. These results suggest that finetuning Whisper-Tiny on the Italian data substantially enhances its recognition capabilities, making it a viable option for tasks requiring higher accuracy. The Whisper-Base model demonstrated a WER of 45% in its baseline form, which was already an improvement compared to performance without the finetuning of Whisper-Tiny. After finetuning, Whisper-Base achieved a WER of 21.7%, the best performance among the models tested. The CER also showed a notable improvement, reaching 7.28%. The finetuning process significantly improves the performance of Whisper models, especially Whisper-Base, which shows the most promising results for high-accuracy transcription tasks.

4. Conclusion and Future Works

In this paper, we conducted a comparative study to understand the best ASR systems for real-time applications on devices with constrained computational resources. We evaluated three widely used ASR systems: Whisher-tiny, Whisher-base, and Vosk API in Italian. Although significant progress has been made in recent years in ASR models, this progress has not been uniformly distributed across all languages, as the availability and quality of datasets heavily influence it. To understand such ASR systems for real-time applications, we fine-tuned the three pre-trained models using a public Italian dataset called the Common Voice dataset. The performances of the three models are comparable to those in the literature for English, the most commonly used language.

In future works, we consider other models, such as WavLM [12], SeamlessM4T [16] and DistilWhisper [17]. Moreover, we also intend to consider other datasets, such as the Italian Spontaneous Dialogue Dataset and Speech DatCar database [18], to study the robustness of such methods in noisy environments. To address the current limitations of spoken Italian datasets, we intend to develop a new dataset that expands demographic and linguistic diversity, including underrepresented dialects, minority languages, and diverse speaker groups. Another challenge is to apply such systems to other scopes, like emotion recognition or stress detections from speech, to integrate different approaches based on analysis of wearable sensors data like proposed in [19, 20] Another challenging aspect is to extend the approach proposed in [21, 22] for programming Graph Neural Networks to transformers and DNN-HMMs in the ASR applicative scenarios.

Acknowledgments

This work has been funded by the European Union - NextGenerationEU under the Italian Ministry of University and Research (MUR) National Innovation Ecosystem grant ECS00000041 - VITALITY - CUP J13C22000430001.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT and Grammarly in order to: Grammar and spelling check, Paraphrase and reword. After using this tool/service, the authors reviewed and edited the content as needed and takes full responsibility for the publication’s content.

References

[1] A. P. Association, et al., Diagnostic and statistical manual of mental disorders, Text revision (2000).

[2] P. D. Emmady, C. Schoo, P. Tadi, Major neurocognitive disorder (dementia) (2020).

[3] C. Reitz, R. Mayeux, Alzheimer disease: epidemiology, diagnostic criteria, risk factors and biomarkers, Biochemical pharmacology 88 (2014) 640–651.

[4] S. S. Aldharman, F. T. Alayed, B. S. Aljohani, A. M. Aladwani, M. A. Alyousef, K. M. Hakami, D. M. Albalawi, S. A. Alnaaim, F. Alayed, A. Aladwani, et al., An assessment of dementia knowledge and its associated factors among health college students in saudi arabia, Cureus 15 (2023).

[5] J. Galvin, C. Roe, K. Powlishta, M. Coats, S. Muich, E. Grant, J. Miller, M. Storandt, J. Morris, The ad8: a brief informant interview to detect dementia, Neurology 65 (2005) 559–564.

[6] C. Burger, M. C. Lopez, J. A. Feller, H. V. Baker, N. Muzyczka, R. J. Mandel, Changes in transcription within the ca1 field of the hippocampus are associated with age-related spatial learning impairments, Neurobiology of learning and memory 87 (2007) 21–41.

[7] 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.

[8] N. V. Shmyrev, other contributors, Vosk speech recognition toolkit: Offline speech recognition api for android, ios, raspberry pi and servers with python, java, c# and node, https://github.com/alphacep/vosk-api, 2020. Accessed: 2025-05-16.

[9] M. Malik, M. K. Malik, K. Mehmood, I. Makhdoom, Automatic speech recognition: a survey, Multimedia Tools and Applications 80 (2021) 9411–9457.

[10] S. Schneider, A. Baevski, R. Collobert, M. Auli, wav2vec: Unsupervised pre-training for speech recognition, arXiv preprint arXiv:1904.05862 (2019).

[11] T. A. Nguyen, B. Sagot, E. Dupoux, Are discrete units necessary for spoken language modeling?, IEEE Journal of Selected Topics in Signal Processing 16 (2022) 1415–1423.

[12] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao, et al., Wavlm: Large-scale self-supervised pre-training for full stack speech processing, IEEE Journal of Selected Topics in Signal Processing 16 (2022) 1505–1518.

[13] I. Dataset, Defined.ai., https://www.defined.ai/datasets/italian-spontaneous-dialogue, ???? Accessed: 2025-05-16.

[14] Common voice mozilla. accessed: Feb. 2, 2025., 2017.

[15] A. C. Morris, V. Maier, P. D. Green, From wer and ril to mer and wil: improved evaluation measures for connected speech recognition., in: Interspeech, 2004, pp. 2765–2768.

[16] L. Barrault, Y.-A. Chung, M. C. Meglioli, D. Dale, N. Dong, P.-A. Duquenne, H. Elsahar, H. Gong, K. Heffernan, J. Hoffman, et al., Seamlessm4t: Massively multilingual & multimodal machine translation, arXiv preprint arXiv:2308.11596 (2023).

[17] S. Gandhi, P. von Platen, A. M. Rush, Distil-whisper: Robust knowledge distillation via large-scale pseudo labelling, arXiv preprint arXiv:2311.00430 (2023).

[18] ELRA, Catalogue, Italian speechdat-car database – elra catalogue, https://github.com/alphacep/vosk-api, 2017. Accessed: 2025-05-16.

[19] M. Quadrini, A. Capuccio, D. Falcone, S. Daberdaku, A. Blanda, L. Bellanova, G. Gerard, Stress detection with encoding physiological signals and convolutional neural network, Machine Learning 113 (2024) 5655–5683.

[20] M. Serenelli, M. Quadrini, M. Óskarsdóttir, M. Loreti, Encoding methods comparison for stress detection, in: CEUR WORKSHOP PROCEEDINGS, volume 3762, CEUR-WS, 2024, 2022.

[21] M. Belenchia, F. Corradini, M. Quadrini, M. Loreti, libmg: A python library for programming graph neural networks in μg, Science of Computer Programming 238 (2024) 103165.

[22] M. Belenchia, F. Corradini, M. Quadrini, M. Loreti, Implementing a ctl model checker with μ g, a language for programming graph neural networks, in: International Conference on Formal Techniques for Distributed Objects, Components, and Systems, Springer, 2023, pp. 37–54.