ASR Systems Under Acoustic Challenges: A Multilingual Study

Sergei Katkov[1][0009−0006−1000−295X], Antonio Liotta[1][0000−0002−2773−4421], and Alessandro Vietti[1][0000−0002−4166−540X]

Free University of Bozen-Bolzano, Bolzano 39100, Italy

Abstract. The performance of automatic speech recognition (ASR) systems in acoustically challenging environments is crucial for the effectiveness of various voice-controlled applications. This study presents an extensive experimental evaluation of the robustness of different ASR models against a range of acoustic disturbances, including white noise, reverberation, time stretch, and pitch shift. By comparing the performance of these models in English, Italian, and German, this research provides a cross-linguistic perspective. The findings reveal a significant decline in performance across all models when subjected to these audio distortions, highlighting the varying degrees of resilience across different languages. By incorporating multiple languages, this study offers valuable insights into the unique challenges and potential opportunities for enhancing ASR technologies, addressing both well-researched and less-explored linguistic domains. Our comparative study highlights that although ASRs are reaching near-human accuracy in ideal acoustic conditions, ASR performance under the whole range of distortions is still well below human performance

Keywords: Automatic Speech Recognition

1 Introduction

Recent advancements in automatic speech recognition (ASR) have introduced models like Whisper [28], Conformer [13], and QuartzNet [19]. These developments have significantly enhanced ASR efficiency and speed, which is essential for a wide range of applications.

However, the accuracy of ASR models, especially under acoustically challenging conditions, remains crucial. Prior studies [21,14,5] have highlighted the importance of improving ASR systems' resilience to noise and other acoustic distortions. Improving ASR robustness to noise is a direct approach to enhancing performance, particularly under conditions with significant noise.

This research evaluates the robustness of ASR models, including the Whisper family, QuartzNet, Conformer, and Fast Conformer, under various audio transformations — such as white noise, reverberation, time stretching, and pitch shifting — chosen to replicate common auditory challenges encountered in real-life and online communication scenarios. These transformations are selected to mimic common auditory challenges encountered in real life and online communication.

Our study seeks to contribute to the ASR field by systematically investigating various ASR models under diverse acoustic conditions and across multiple languages. The findings are intended to inform future developments in speech recognition technology, optimizing its application in varied real-world scenarios and expanding its utility across different linguistic domains.

2 Related Work

Advances in automatic speech recognition (ASR) technology, especially with models like Whisper [28], Conformer [13], and QuartzNet [19], have been significant.

The native multilingual capabilities of models like Whisper [28], compared to fine-tuning methods [16], represent different strategies for adapting ASR technologies across languages and conditions, enhancing their generalizability and utility.

While the Whisper model shows resilience in basic noise environments [28,22], its performance under extensive acoustic variations remains less explored. Conformer's integration into denoising pipelines [7,20] showcases improvements in recognition amidst noise. Developing noisy datasets [6] and noise augmentation techniques [1] has been essential, though their applicability to Whisper and QuartzNet requires more exploration.

In [18], it is shown that ASR models degrade significantly at high noise levels in Italian, even when human listeners can transcribe accurately. In [17], Whisper models struggle with audio transformations and chunk length variations, particularly for German. However, the study lacks a multilingual comparison, highlighting the need for broader cross-linguistic analysis.

Research on noise removal [32,21] and speech dereverberation [30,31] offers solutions to mitigate auditory distortions, which are common challenges in ASR applications. These studies lay a foundation for enhancing ASR robustness to noise and reverberation.

QuartzNet, when fine-tuned with noise augmentations, shows improvements in handling noisy samples while maintaining performance on clean data [3], demonstrating the potential of targeted noise augmentation.

Pitch manipulation research aims to reduce performance gaps between male and female voices [9], a critical area for ensuring ASR systems handle cross-speaker variation effectively.

In summary, ASR has progressed significantly with models like Whisper, Conformer, and QuartzNet, but further exploration is needed in noise handling, unconventional transformations, and multilingual support. These areas offer promising paths for enhancing ASR robustness and versatility.

Our research evaluates the robustness of Whisper, QuartzNet, Conformer, and Fast Conformer models under diverse acoustic disturbances in English, Italian, and German languages.

3 Methodology

This study assesses Whisper, QuartzNet, Conformer, and Fast Conformer ASR models' robustness to audio disturbances, focusing on English, Italian, and German languages. We conduct transformations to mimic challenges encountered in online communications and real-world environments, evaluating their performance and identifying areas for enhancement.

3.1 Models

The Whisper, QuartzNet, Conformer, and Fast Conformer models were selected for their architectural characteristics and their different approaches to handling multilingual data and noise.

Whisper Models We utilized the Whisper base, medium, and large-v3 models [28], leveraging their multilingual capabilities by specifying the target language (English, German, or Italian) during inference. These models are designed to support multiple languages, allowing for optional language selection at inference.

QuartzNet 15x5 QuartzNet [19] 15x5, featuring a deep 79-layer architecture and 18.9 million parameters, was initially pretrained on English datasets such as LibriSpeech [24], Fisher Corpus [4], Switchboard-1 [10], WSJ-0, and WSJ-1 [25]. It was subsequently fine-tuned for various languages using the Common Voice [2] dataset.

Conformer CTC Large The Conformer CTC Large model, which uses around 120 million parameters, employs the Connectionist Temporal Classification (CTC) loss function [12]. It was trained on datasets such as Common Voice [2], Multilingual LibriSpeech [27], and VoxPopuli [33], and utilizes a SentencePiece tokenizer [11].

Conformer-Transducer Large This model utilizes the Recurrent Neural Network Transducer (RNNT) loss and decoder [34] for automatic speech recognition. It was trained on the same datasets as the Conformer CTC Large.

FastConformer Hybrid Transducer-CTC Large The FastConformer [29] Hybrid Transducer-CTC model combines the strengths of both CTC and Transducer models. It was trained on the same speech data as the Conformer models. The architecture of this model is optimized with 8x depthwise-separable convolutional downsampling.

Table 1 provides a summary of the model architectures and parameter counts for each ASR model evaluated in this study.

Model	Architecture	Parameters (millions)
Whisper Base	Transformer	74
Whisper Medium	Transformer	769
Whisper Large-v3	Transformer	1550
QuartzNet 15x5	CNN	18.9
Conformer-CTC Large	Conformer	120
Conformer-Transducer Large	Conformer	120
FastConformer Hybrid	Conformer	114

Table 1. Model architectures and parameter counts

3.2 Dataset

To evaluate the efficiency of the ASR models in environments augmented with audio disturbances, we utilized the test subsets of the Common Voice 13.0 dataset [2] for English, Italian, and German. The dataset consists of approximately 13,000 utterances per language, providing a balanced representation of various accents and speech contexts encountered in real-world scenarios.

3.3 Evaluation Metrics

To assess the performance of speech recognition systems in our study, we employ the Word Error Rate (WER) metric. The WER is calculated as follows:

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

where S, D, and I denote the numbers of substitutions, deletions, and insertions needed to match the system's transcription to the reference text. N represents the total count of words in the reference text.

This measure serves as an indicator of transcription accuracy, with lower WER values reflecting better performance. Typically, a WER below 0.1 is considered excellent, 0.1-0.2 is acceptable but may indicate potential issues, and above 0.2 denotes significant transcription errors, making the ASR output difficult to understand.

For text normalization, punctuation and other non-alphanumeric symbols were removed, and all text was converted to lowercase.

3.4 Audio Transformations

To evaluate the performance of ASR models under realistic acoustic conditions, specific audio transformations were applied. These transformations were chosen to replicate common auditory challenges.

The white noise transformation adds uniform noise across various frequencies, simulating background noise found in crowded places, urban settings, and telecommunications or online communications due to signal interference or compression artifacts. Time stretch transformation changes the duration of an audio signal without altering its pitch, mimicking scenarios where speech speed varies, such as in spontaneous conversations. Pitch shift transformation changes the pitch of an audio signal, representing different speaker fundamental frequency, singing voices, and speech patterns of individuals with certain medical conditions, testing the model's adaptability to varying vocal pitches. Reverberation adds echo effects to simulate environments like large halls, reflective rooms, and phone calls, testing the model's ability to handle echoes.

The transformations are defined as follows, inspired by real-world auditory conditions to evaluate model robustness:

- **White Noise:** A uniform noise signal added across various frequencies to simulate background noise in urban settings or online communications, expressed as
$$n(t) = \alpha \cdot \mathrm{rand}(t), \tag{2}$$
where α represents the amplitude.
- **Time Stretch:** Modifies the duration of an audio signal without altering its pitch, representing variations in speech speed in conversations, described by
$$y(t) = x(a \cdot t), \tag{3}$$
where a is the stretch factor.
- **Pitch Shift:** Alters the pitch using Fourier Transform techniques, reflecting different vocal pitches, given by
$$y(t) = F^{-1}\{F\{x(t)\} \cdot e^{j2\pi \Delta f t}\}, \tag{4}$$
with Δf indicating the frequency shift.
- **Reverberation:** Simulates echo effects as in large rooms or phone calls, represented as
$$y(t) = x(t) + \alpha \cdot x(t - \Delta t), \tag{5}$$
where α is the decay rate and Δt is the delay time.

4 Results

We analyzed Whisper, QuartzNet, and Conformer models under acoustic disturbances like white noise, time stretch, pitch shift and reverberation to explore their robustness in the English, Italian and German context.

The results in Table 2 show that the Whisper Large model and Conformer variants consistently outperform other models, while QuartzNet's lower accuracy reflects its simpler architecture. The Whisper Base model, despite being the smallest in the advanced Whisper series, has the highest WER, highlighting its limitations in achieving optimal accuracy as a compact transformer model. This result underscores the trade-off between model complexity and performance, particularly in noise-free environments. Across different languages, this tendency persists. While the Whisper Base model excels in English, the more advanced

Model / Language	English	Italian	German
Whisper Base	0.26	0.37	0.30
Whisper Medium	0.13	0.10	0.09
Whisper Large-v3	0.11	0.06	0.06
QuartzNet	0.22	0.17	0.14
Conformer-CTC Large	0.10	0.07	0.07
Conformer-Transducer Large	0.08	0.05	0.06
FastConformer-Hybrid CTC/Transducer	0.10	0.06	0.05

Table 2. WER for ASR Models in Noise-Free Scenario

Whisper Large-v3 model performs better in Italian and German. The superior performance of QuartzNet and Conformer models can be attributed to the use of the Common Voice dataset for tuning non-English languages, enabling these models to better adapt to the distribution in the test set.

4.1 English Language Experiments

We performed a set of experiments on all previously listed models for the English language.

The almost linear degradation in quality is evident for nearly all models, as shown in Fig. 1. The Whisper Base model appears to be the least robust, with its WER increasing more rapidly than that of other models. Among the other models, QuartzNet demonstrates poorer performance compared to the rest. At a noise level of 0.03, all models experience a marked reduction in quality, highlighting a clear deviation from the human ability to comprehend and interpret audio content in similar conditions [26].

Although pitch changes do not greatly affect human comprehension of audio, these modifications result in a noticeable and fairly consistent decline in ASR model performance across all tested levels. Of the models evaluated, the Conformer Transducer Large shows the greatest resilience to pitch alterations, as indicated in Fig. 2. It is important to note that the specific level of pitch variation has minimal impact on the model. It seems that merely shifting the signal out of the training set distribution is sufficient to significantly degrade performance.

At altered time stretch levels, there is a universal decline in performance across all models, as shown in Fig. 3. The x-axis value of 1.0 represents no transformation, with values to the left indicating slowed down audio and values to the right indicating sped up audio. However, the Whisper models, particularly the smaller variants, experience a more pronounced performance drop and are prone to generating repetitive phrases in their outputs. This phenomenon, known as "hallucination" is widely observed in sequence generation models and affects both ASR [8] and broader language generation [15], leading to significantly inflated WER. Notably, even at stretch rates of 0.9 or 1.1, where humans find the audio completely intelligible [23], there is still a noticeable decline in recognition

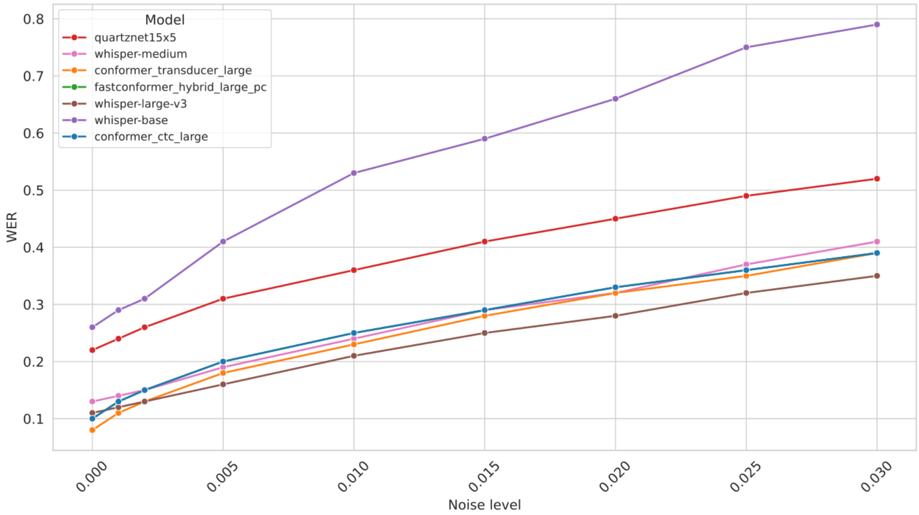

Fig. 1. WER comparison for different models under white noise for English language

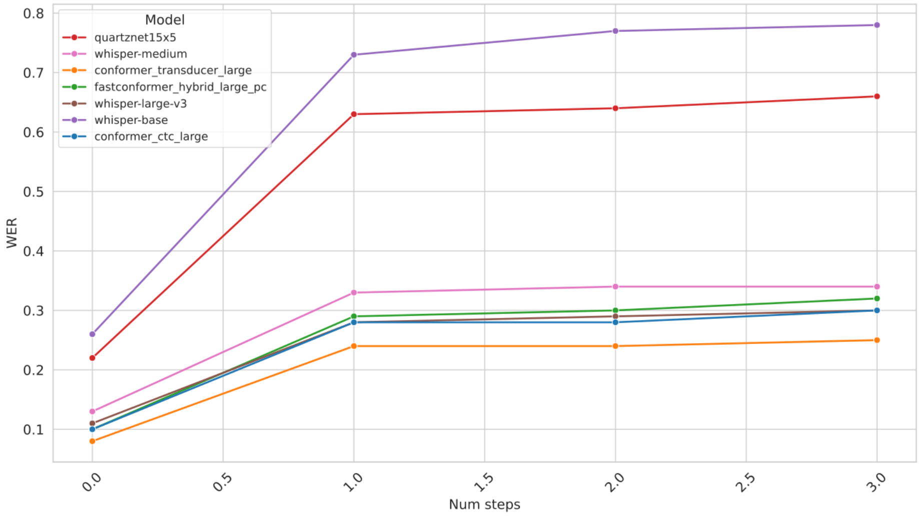

Fig. 2. WER comparison for different models under pitch shift for English language

accuracy for these models. Interestingly, the Conformer Transducer Large performs better with sped-up audio, while Whisper Large-v3 handles slowed-down audio more effectively.

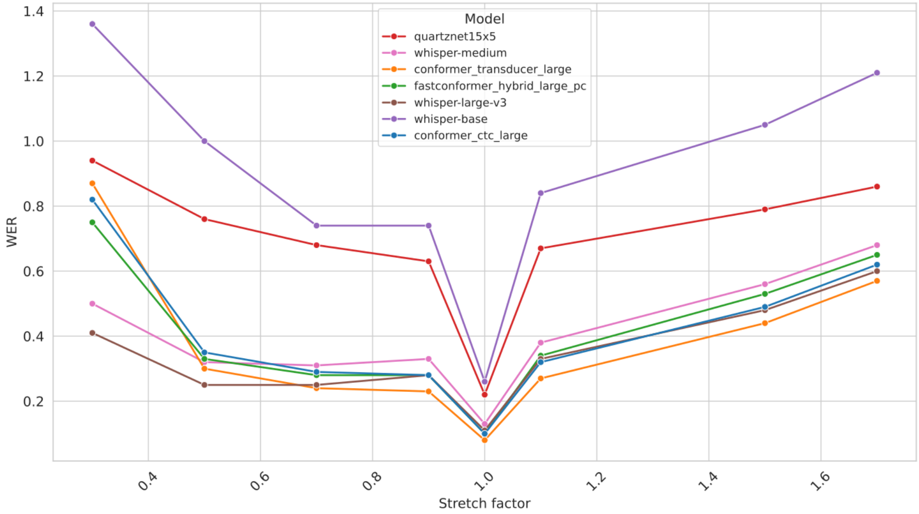

Fig. 3. WER comparison for different models under time stretch for English language

Fig. 4 illustrates the superior performance the Whisper Large-v3 model, in handling reverberated audio. In contrast, the Whisper Base model struggles significantly with these transformations. Notably, across all tested reverberation times, we observe a nearly uniform degradation in model performance, indicating that these ASR systems are sensitive to the presence of reverberation rather than its intensity.

4.2 Multilingual Experiments

Experiments were conducted for English, Italian, and German languages. To maintain concise and clear visual representations, we present the results only for the Whisper Large-v3, Whisper Base, and Conformer Transducer Large models. For each language, a distinct color is used, and for each model, a unique line style is applied. Specifically, the Whisper Large-v3 model is represented with a solid line, the Conformer Transducer Large with dashed lines, and the Whisper Base with dotted lines.

The Conformer model demonstrates superior robustness in Italian and German, while the Whisper Large-v3 model performs better in English. As shown in Fig. 5, both the models achieve commendable results, with their performance in Italian and German surpassing that in English. This may be attributed to the phonetic properties of these languages. The Whisper Base model, however,

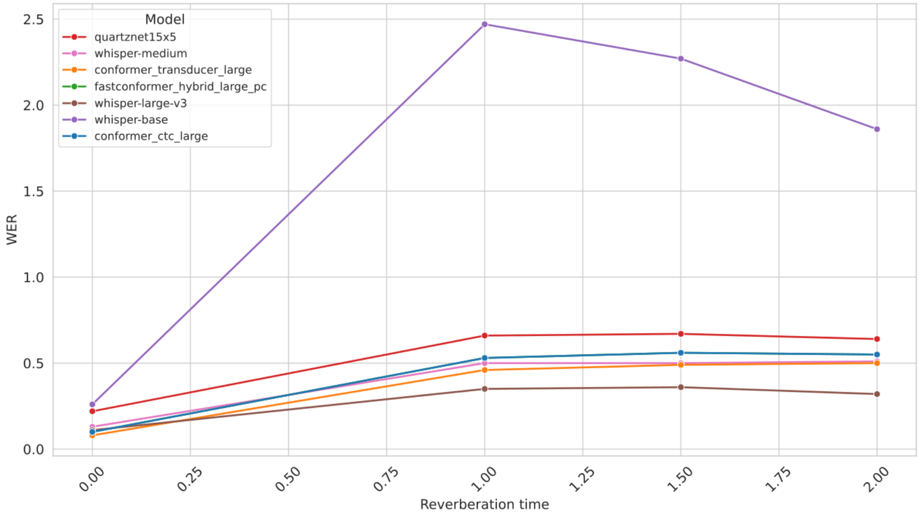

Fig. 4. WER comparison for different models under reverberation for English language

exhibits significantly lower quality in Italian compared to other languages. Conversely, for English, the Whisper Base model shows relatively good performance, whereas the Whisper Large-v3 and Conformer models have higher WER.

The same tendency persists for pitch shift transformations. As shown in Fig. 6, the Whisper Large-v3 model performs almost equally well for both German and Italian languages. The Conformer model, however, demonstrates slightly better performance in the Italian language.

For time stretch transformations, the Whisper Large-v3 and Conformer models perform at a similar quality overall. However, the Conformer model performs better for stretch factors greater than 1.0, while the Whisper Large-v3 model performs better for stretch factors less than 1.0, as shown in Fig. 7.

It can be observed that the Whisper Large-v3 model is the most robust to reverberation across all languages. Although the WER remains relatively high, it is significantly lower compared to other models, as illustrated in Fig. 4.

The performance differences across languages can be attributed to linguistic features. Italian and German have more consistent phoneme-to-grapheme mappings than English, possibly explaining the better model performance. Additionally, German compound words and Italian vowel-rich phonetics pose unique challenges. Future work could explore these nuances further to optimize ASR model training and fine-tuning for specific languages.

Overall, the Whisper Large-v3 model is more robust to transformations such as reverberation and time stretch with a stretch factor less than 1.0 (slowing down). In contrast, the Conformer model performs better under conditions such as white noise. This robustness can be attributed to the specific fine-tuning processes applied to these models.

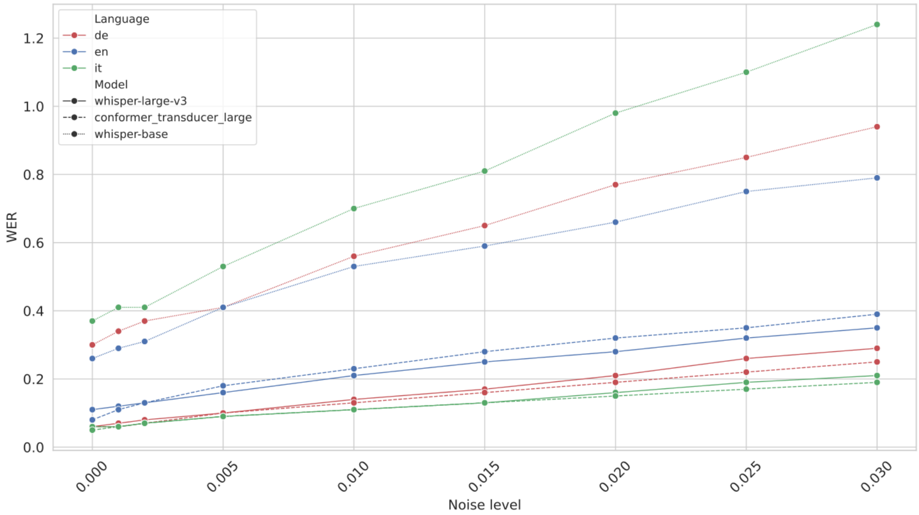

Fig. 5. WER comparison for different models and languages under white noise

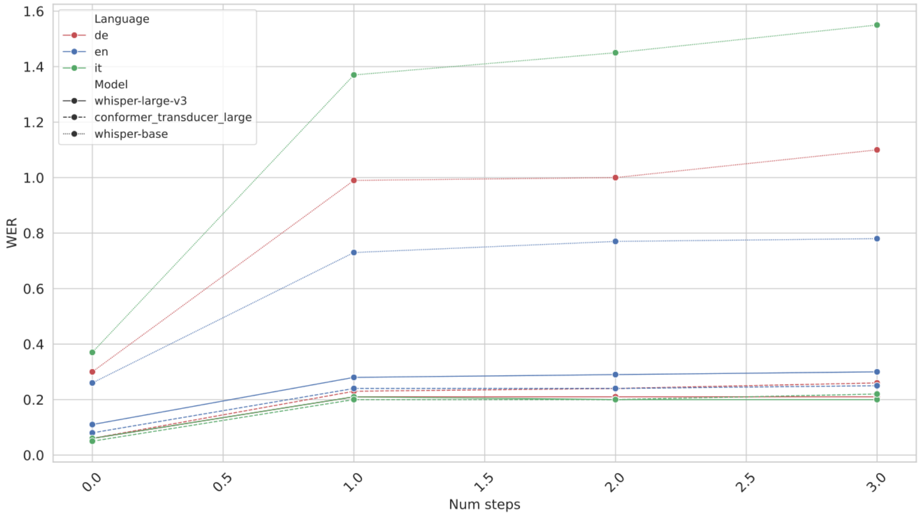

Fig. 6. WER comparison for different models and languages under pitch shift

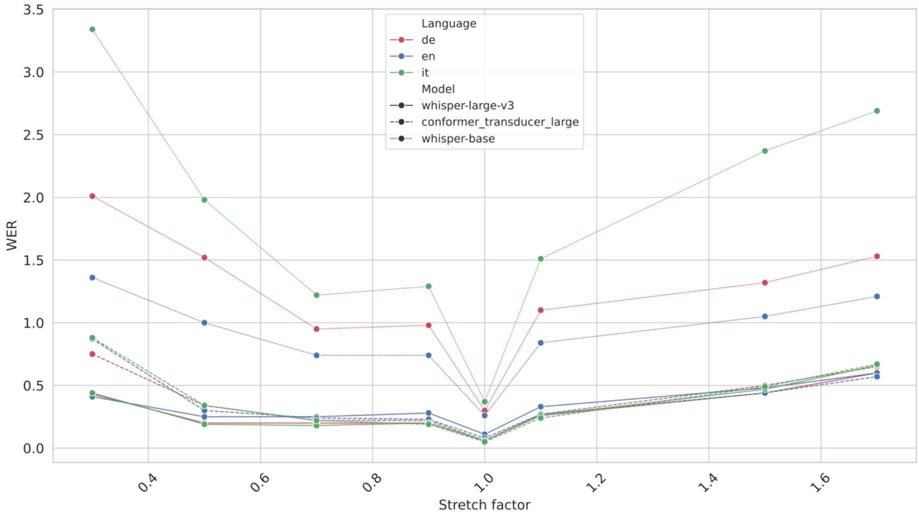

Fig. 7. WER comparison for different models and languages under time stretch

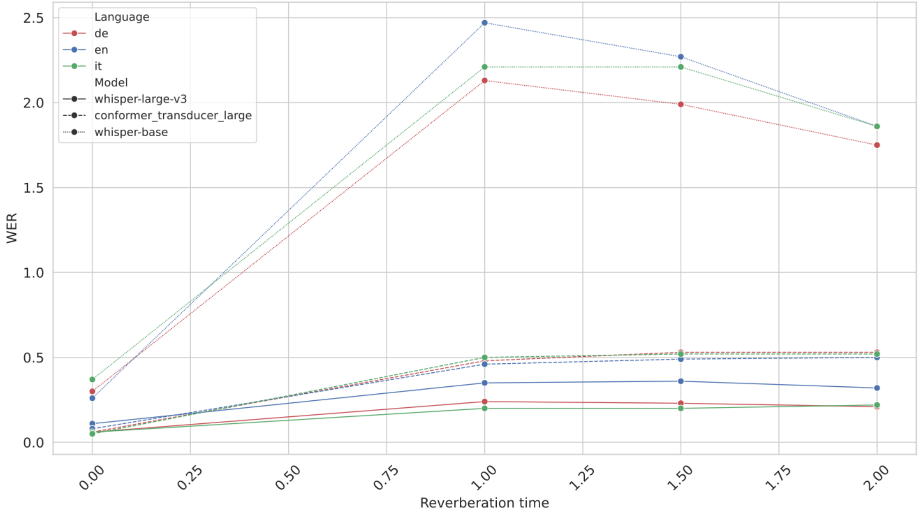

Fig. 8. WER comparison for different models under reverberation

5 Conclusion

This study evaluates Whisper, QuartzNet, and Conformer ASR models against acoustic disturbances (white noise, reverberation, time stretch, pitch shift) across English, Italian, and German. The findings highlight each model's unique response to these challenges, with larger models like Whisper Large-v3 and Conformer generally performing better, though they still struggle with certain transformations. Whisper Base, with its limited parameters, exhibits significant robustness issues and a tendency to hallucinate.

Interestingly, despite more training data for English, Italian and German often show better ASR performance, suggesting language-specific factors in ASR accuracy. Whisper's multilingual capability is notable, but it sometimes underperforms compared to specialized Conformer models, indicating a trade-off between versatility and accuracy.

Different audio transformations uniquely affect ASR model performance. For example, reverberation shows a uniform degradation across models, suggesting a need for specialized training to handle such transformations better.

While ASR achieves near-human accuracy under ideal conditions, its performance under distortions is still below human levels. Noises that humans can easily compensate for result in high WER for ASR systems. For instance, white noise often produces high WER despite humans understanding the speech relatively well. Similarly, time stretch transformations might not significantly hinder human comprehension but can drastically increase WER for ASR models.

Future research should explore advanced noise augmentation techniques, understand linguistic nuances contributing to performance differences, and ensure balanced language representation in training data. Additionally, the acoustic transformations used in this study could provide insights for improving ASR systems for pathological speech, which often exhibits irregular pitch, breathiness, and other distortions. Addressing these aspects will make ASR systems more inclusive and capable of serving users with a wide range of speech characteristics. Furthermore, investigating the impact of mixed noise scenarios—where multiple types of acoustic disturbances, such as background noise and reverberation, occur simultaneously — can offer a more comprehensive understanding of ASR robustness in complex real-world environments, guiding the development of more resilient systems.

It is also essential to investigate the discrepancies between human and machine intelligibility of distorted speech, aiming to develop ASR systems that align more closely with human auditory perception. These steps will help develop ASR technologies that are robust, reliable, and effective across diverse linguistic contexts.

Acknowledgments. This research was conducted as part of the RATTLE (Voice Recogniser based on Artificial Intelligence) project, kindly funded by the Fondazione Pfizer.

References

1. Adolfi, F., Bowers, J.S., Poeppel, D.: Successes and critical failures of neural networks in capturing human-like speech recognition. Neural Netw. **162**(C), 199–211 (may 2023). https://doi.org/10.1016/j.neunet.2023.02.032, https://doi.org/10.1016/j.neunet.2023.02.032
2. Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F.M., Weber, G.: Common voice: A massively-multilingual speech corpus. In: International Conference on Language Resources and Evaluation (2019), https://api.semanticscholar.org/CorpusID:209376338
3. Balam, J., Huang, J., Lavrukhin, V., Deng, S., Majumdar, S., Ginsburg, B.: Improving noise robustness of an end-to-end neural model for automatic speech recognition (2020)
4. Cieri, C., Miller, D., Walker, K.: The fisher corpus: A resource for the next generations of speech-to-text (01 2004)
5. Cui, T., Xiao, J., Li, L., Jiang, X., Liu, Q.: An approach to improve robustness of nlp systems against asr errors. ArXiv **abs/2103.13610** (2021), https://api.semanticscholar.org/CorpusID:232352551
6. Duarte, J.C., Colcher, S.: Building a noisy audio dataset to evaluate machine learning approaches for automatic speech recognition systems. ArXiv **abs/2110.01425** (2018), https://api.semanticscholar.org/CorpusID:238259030
7. Eickhoff, P., Möller, M., Pekarek-Rosin, T., Twiefel, J., Wermter, S.: Bring the noise: Introducing noise robustness to pretrained automatic speech recognition. In: International Conference on Artificial Neural Networks (2023), https://api.semanticscholar.org/CorpusID:261559431
8. Frieske, R., Shi, B.E.: Hallucinations in neural automatic speech recognition: Identifying errors and hallucinatory models (2024)
9. Fucci, D., Gaido, M., Negri, M., Cettolo, M., Bentivogli, L.: No pitch left behind: Addressing gender unbalance in automatic speech recognition through pitch manipulation. 2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU) pp. 1–8 (2023), https://api.semanticscholar.org/CorpusID:263830339
10. Godfrey, J., Holliman, E., McDaniel, J.: Switchboard: telephone speech corpus for research and development. In: [Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing. vol. 1, pp. 517–520 vol.1 (1992). https://doi.org/10.1109/ICASSP.1992.225858
11. Google: Sentencepiece. https://github.com/google/sentencepiece
12. Graves, A., Fernández, S., Gomez, F., Schmidhuber, J.: Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural 'networks. vol. 2006, pp. 369–376 (01 2006). https://doi.org/10.1145/1143844.1143891
13. Gulati, A., Chiu, C.C., Qin, J., Yu, J., Parmar, N., Pang, R., Wang, S., Han, W., Wu, Y., Zhang, Y., Zhang, Z. (eds.): Conformer: Convolution-augmented Transformer for Speech Recognition (2020)
14. Higuchi, Y., Tawara, N., Ogawa, A., Iwata, T., Kobayashi, T., Ogawa, T.: Noise-robust attention learning for end-to-end speech recognition. In: 2020 28th European Signal Processing Conference (EUSIPCO). pp. 311–315 (2021). https://doi.org/10.23919/Eusipco47968.2020.9287488
15. Holtzman, A., Buys, J., Forbes, M., Choi, Y.: The curious case of neural text degeneration. CoRR **abs/1904.09751** (2019), http://arxiv.org/abs/1904.09751
16. Huang, J., Kuchaiev, O., O'Neill, P., Lavrukhin, V., Li, J., Flores, A., Kucsko, G., Ginsburg, B.: Cross-language transfer learning, continuous learning, and domain adaptation for end-to-end automatic speech recognition (2020)

17. Katkov, S., Liotta, A., Vietti, A.: Benchmarking whisper under diverse audio transformations and real-time constraints. In: Proceedings of the 26th International Conference on Speech and Computer (SPECOM) (2024)
18. Katkov, S., Liotta, A., Vietti, A.: Evaluating the robustness of ASR systems in adverse acoustic conditions. In: Proceedings of the Fifth International Conference on Intelligent Data Science Technologies and Applications (IDSTA) (2024)
19. Kriman, S., Beliaev, S., Ginsburg, B., Huang, J., Kuchaiev, O., Lavrukhin, V., Leary, R., Li, J., Zhang, Y.: Quartznet: Deep automatic speech recognition with 1d time-channel separable convolutions. In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6124–6128 (2020). https://doi.org/10.1109/ICASSP40776.2020.9053889
20. Lee, G.W., Kim, H.K.: Two-step joint optimization with auxiliary loss function for noise-robust speech recognition. Sensors (Basel, Switzerland) **22** (2022), https://api.semanticscholar.org/CorpusID:250942334
21. Li, J., Deng, L., Gong, Y., Häb-Umbach, R.: An overview of noise-robust automatic speech recognition. IEEE/ACM Transactions on Audio, Speech, and Language Processing **22**, 745–777 (2014), https://api.semanticscholar.org/CorpusID:14557362
22. Mauch, M., Ewert, S.: The audio degradation toolbox and its application to robustness evaluation. In: International Society for Music Information Retrieval Conference (2013), https://api.semanticscholar.org/CorpusID:11675708
23. Müller, J.A., Wendt, D., Kollmeier, B., Debener, S., Brand, T.: Effect of speech rate on neural tracking of speech. Frontiers in Psychology **10** (2019). https://doi.org/10.3389/fpsyg.2019.00449, https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2019.00449
24. Panayotov, V., Chen, G., Povey, D., Khudanpur, S.: Librispeech: An asr corpus based on public domain audio books. In: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 5206–5210 (2015). https://doi.org/10.1109/ICASSP.2015.7178964
25. Paul, D.B., Baker, J.M.: The design for the wall street journal-based csr corpus. In: Proceedings of the Workshop on Speech and Natural Language. p. 357–362. HLT '91, Association for Computational Linguistics, USA (1992). https://doi.org/10.3115/1075527.1075614, https://doi.org/10.3115/1075527.1075614
26. Payton, K.L., Uchanski, R.M., Braida, L.D.: Intelligibility of conversational and clear speech in noise and reverberation for listeners with normal and impaired hearing. The Journal of the Acoustical Society of America **95**(3), 1581–1592 (03 1994). https://doi.org/10.1121/1.408545, https://doi.org/10.1121/1.408545
27. Pratap, V., Xu, Q., Sriram, A., Synnaeve, G., Collobert, R.: Mls: A large-scale multilingual dataset for speech research. In: Interspeech 2020. ISCA (Oct 2020). https://doi.org/10.21437/interspeech.2020-2826, http://dx.doi.org/10.21437/Interspeech.2020-2826
28. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision (2022)
29. Rekesh, D., Koluguri, N.R., Kriman, S., Majumdar, S., Noroozi, V., Huang, H., Hrinchuk, O., Puvvada, K., Kumar, A., Balam, J., Ginsburg, B.: Fast conformer with linearly scalable attention for efficient speech recognition (2023)
30. Saito, K., Murata, N., Uesaka, T., Lai, C.H., Takida, Y., Fukui, T., Mitsufuji, Y.: Unsupervised vocal dereverberation with diffusion-based generative models (2022)
31. Schwartz, B., Gannot, S., Habets, E.: Online speech dereverberation using kalman filter and em algorithm. IEEE/ACM Transactions on Audio, Speech, and Language Processing **23**, 394–406 (2015), https://api.semanticscholar.org/CorpusID:2413399

32. Shrawankar, U., Thakare, V.: Noise estimation and noise removal techniques for speech recognition in adverse environment. In: Shi, Z., Vadera, S., Aamodt, A., Leake, D. (eds.) Intelligent Information Processing V. pp. 336–342. Springer Berlin Heidelberg, Berlin, Heidelberg (2010)
33. Wang, C., Rivière, M., Lee, A., Wu, A., Talnikar, C., Haziza, D., Williamson, M., Pino, J.M., Dupoux, E.: Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. CoRR **abs/2101.00390** (2021), https://arxiv.org/abs/2101.00390
34. Zhang, Q., Lu, H., Sak, H., Tripathi, A., McDermott, E., Koo, S., Kumar, S.: Transformer transducer: A streamable speech recognition model with transformer encoders and rnn-t loss. In: ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 7829–7833 (2020). https://doi.org/10.1109/ICASSP40776.2020.9053896