Evaluating the Robustness of ASR Systems in Adverse Acoustic Conditions

Sergei Katkov*, Antonio Liotta*, Alessandro Vietti*
*Free University of Bozen-Bolzano, Bolzano, Italy

***Abstract*—The effectiveness of automatic speech recognition (ASR) systems in environments with acoustic challenges directly influences their utility in a range of voice-activated applications. This paper focuses on an experimental analysis of the resilience of various ASR models to acoustic disturbances — specifically white noise, reverberation, time stretch, and pitch shift — within the context of the Italian language, a non-English and comparatively less-studied linguistic domain. The investigation reveals a notable degradation in performance across the board when models are subjected to these audio transformations. By focusing on Italian, this research contributes valuable insights into the challenges and opportunities in optimizing ASR technologies for languages with lower research exposure.**

***Index Terms*—Automatic Speech Recognition, Deep Learning**

I. Introduction

Recent advancements in automatic speech recognition (ASR) have introduced models like wav2vec 2.0 [1], Whisper [2], and Conformer [3], along with compact models like Jasper [4] and QuartzNet [5]. These developments have significantly enhanced ASR efficiency and speed, which are essential for a wide range of applications.

However, the accuracy of ASR models, especially under acoustically challenging conditions, remains crucial. Prior studies [6]–[8] have highlighted the importance of improving ASR systems' resilience to noise and other auditory distortions.

This research evaluates the robustness of ASR models against white noise, reverberation, time stretching, and pitch shifting, specifically focusing on the Italian language. This focus is particularly novel as it addresses a less-explored linguistic domain in ASR research. By examining these models' performance under varied acoustic disturbances, we aim to contribute to enhancing ASR technology's adaptability and reliability across languages, particularly in real-world environments where such disturbances are common.

II. Related Work

The field of automatic speech recognition (ASR) has seen significant progress, particularly with the introduction of models like Whisper [2], Conformer [3], and QuartzNet [5], have been crucial.

The native multilingual capabilities of models like Whisper [2], as opposed to the fine-tuning methods [9] for languages like Italian, reflect different strategies for adapting ASR technologies to various languages.

While the Whisper model demonstrated resilience in basic noise environments [2], [10], its performance under extensive acoustic variations remains less explored. Conformer's integration into denoising ASR pipelines [11], [12] showcases innovative approaches to improving speech recognition accuracy amidst noise. The development of noisy datasets [13] and noise augmentation techniques [14] have been essential in enhancing the noise resilience of ASR models. Nonetheless, their applicability to specific models has not been fully explored.

Research on noise removal [6], [15] and speech dereverberation [16], [17] offer various solutions to combat auditory distortions.

QuartzNet, when fine-tuned with noise augmentations, shows notable improvements in handling noisy samples while maintaining performance on clean data [18].

Research on pitch manipulation has been conducted to reduce the performance gap between male and female voices [19], pointing to a significant area for ongoing study.

In summary, the field of ASR has made significant progress with the introduction and refinement of the above models. There are still significant areas for future research, particularly in investigating novel noise conditions and atypical audio transformations, as well as extending support to a broader range of languages. These areas offer promising paths for future studies.

III. Methodology

This study assesses Whisper, QuartzNet, and Conformer ASR models' robustness to audio disturbances focusing on the Italian language. We conduct transformations, to mimic challenges encountered in online communications and real-world environments, to evaluate their performance and identify enhancement areas.

A. Models

The Whisper, QuartzNet, and Conformer models were selected due to their architectural diversity and the availability of versions specifically designed for the Italian language. For experiments we utilize the Whisper model from [20] and Italian QuartzNet, and Conformer models from [21]

a) Whisper Models: We employ Whisper base, medium, and large-v3 models [2], utilizing their multilingual capabilities by specifying Italian as the inference language. These models are designed to be multilingual, supporting optional

This work has been carried out in the context of project RATTLE (Voice Recogniser based on Artificial Intelligence), kindly sponsored by Fondazione Pfizer.

979-8-3503-5475-1/24/$31.00 ©2024 IEEE

language selection at inference. Compared to other tested models, the Whisper variants are notably larger, with the base, medium, and large-v3 models containing 74 million, 769 million, and 1550 million parameters, respectively.

b) QuartzNet 15x5: QuartzNet [5] 15x5, with its deep 79-layer architecture and 18.9 million parameters, originally pretrained on English datasets such as LibriSpeech [22], Fisher Corpus [23], Switchboard-1 [24], WSJ-0, and WSJ-1 [25], with following fine-tuning for Italian with Common Voice 6.0 [26] dataset.

c) Conformer CTC Large It: Conformer CTC Large, leveraging around 120 million parameters, employs the Connectionist Temporal Classification(CTC) loss function [27]. Trained from scratch on a composite dataset of approximately 500 hours of Italian speech, including Common Voice 11.0 [26], Multilingual LibriSpeech [28], and VoxPopuli [29], it utilizes a SentencePiece tokenizer [30] with a vocabulary of 128 tokens.

d) Conformer-Transducer Large It: This model employs the RNNT/Transducer loss/decoder [31] for ASR. It is trained on the same Italian speech dataset as the Conformer CTC Large but uses a tokenizer with a vocabulary size of 1024.

e) FastConformer Hybrid Transducer-CTC Large It: FastConformer [32] Hybrid Transducer-CTC, combining the strengths of both CTC and Transducer models. It was trained on the same speech data as conformer models. This model's architecture is optimized with 8x depthwise-separable convolutional downsampling, and it employs a tokenizer with a vocabulary of 512.

B. Dataset

To evaluate the efficiency of the ASR models in environments with audio disturbances, we utilized the Italian test subset of the Common Voice 13.0 dataset [26]. The test subset comprises 15,096 recorded sentences and 3,753 unique participants. This dataset was selected due to its comprehensive representation of the Italian language, encompassing a wide array of accents, age groups, and speech contexts encountered in real-world scenarios. The diversity of this dataset ensures that the models are tested on a varied set of speech samples, enhancing the relevance of the evaluation to practical applications.

C. Audio Transformations

Evaluating ASR models' performance in real-world-like conditions necessitates applying specific audio transformations. These are selected to replicate common auditory challenges. The transformations include:

- **White Noise:** A signal with uniform intensity across various frequencies, formulated as

$$n(t) = \alpha \cdot \text{rand}(t), \tag{1}$$

  where α represents the amplitude, and $\text{rand}(t)$ is a function generating uniform random numbers.

- **Time Stretch:** Alters the length of an audio clip without changing its pitch, described by

$$y(t) = x(a \cdot t), \tag{2}$$

  a being the stretch factor. This equation allows for the adjustment of the audio's playback speed without affecting the sound's pitch or clarity. Thus, it leads to an alteration in the duration of the audio

- **Pitch Shift:** Modifies an audio signal's pitch using Fourier Transform techniques, expressed as

$$y(t) = F^{-1}\{F\{x(t)\} \cdot e^{j2\pi\Delta f t}\}, \tag{3}$$

  with Δf indicating the frequency shift. The experiment adjusts n_steps, which correlates to Δf through $\Delta f = n\_steps \times \frac{f_0}{12}$, f_0 being the base frequency before modification. An adjustment by one n_steps equates to a semitone pitch change.

- **Reverberation:** Mimics the echo effects in audio, represented as

$$y(t) = x(t) + \alpha \cdot x(t - \Delta t), \tag{4}$$

  where α is the echo decay rate and Δt the delay time.

These audio transformations were selected to challenge the audio processing capabilities of the ASR models, ensuring that the sentences remain comprehensible to human listeners despite the introduced noise or distortion.

IV. Results

To evaluate the performance of the speech recognition systems in our study, we use the Word Error Rate (WER) metric. The WER is computed as:

$$\text{WER} = \frac{S + D + I}{N}, \tag{5}$$

where S represents the number of substitutions, D the number of deletions, and I the number of insertions required to align the system's transcription with the reference text. The total number of words in the reference text is denoted by N. WER serves as a measure of transcription accuracy, with lower values indicating higher accuracy and better performance.

For text normalization, we remove punctuation and other non-alphanumeric symbols, and convert all text to lowercase.

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 Italian context.

TABLE I
WER for ASR Models in No Noise Scenario

Model	WER
Whisper Base	0.37
Whisper Medium	0.10
Whisper Large-v3	0.06
QuartzNet	0.17
Conformer-CTC Large	0.07
Conformer-Transducer Large	0.05
FastConformer-Hybrid CTC/Transducer	0.06

The results in Table I indicate that the Whisper Large model and Conformer variants outshine others, with QuartzNet's lower accuracy reflecting its simpler architecture. The Whisper Base model, despite its smaller size within the advanced Whisper series, records the highest WER, underscoring its limitations as a compact transformer model in achieving optimal accuracy. This result emphasizes the trade-off between model complexity and performance, especially in environments without noise.

In the presentation of our results, a color map has been used to visually delineate the WER performance across audio transformations, with the color coding applied independently to each column. This color scale transitions from green to red to represent a gradation of WER values from the lowest to the highest respectively. To facilitate easier comparison, reference WER values for the no-noise or no-transformation scenario are clearly displayed in the first column of each table, separated by a bold line for clear distinction and quick reference. The transformation levels in each table are arranged from left to right, progressing from least to most severe

Notably, Whisper Large v3 and Conformer Transducer Large exhibit superior resilience in significant white noise augmented conditions (Table II). Important to highlight that although QuartzNet demonstrates a higher WER in noise-free conditions, its performance exhibits a relatively modest degradation rate when compared to other models as noise levels increase. At the noise level of 0.03, there is a significant reduction in quality, revealing a clear divergence from the human ability to discern and understand the audio content in analogous situations [33].

Table III shows Whisper models, especially Large v3, superior performance in handling reverberated audio. Whisper Medium also excels in its effective handling of reverberation, outperforming other models, even those that perform better in clear audio conditions. Remarkably, across all tested reverberation times, we observe a nearly uniform degradation in model performance, suggesting these ASR systems are sensitive to the presence of reverberation rather than its intensity.

While pitch alterations do not significantly impact a person's comprehension of audio, such manipulations across all observed levels lead to a noticeable, relatively consistent decline in ASR model performance. Among the tested models, the Whisper Large v3 and the Conformer Transducer Large demonstrate the strongest resilience to pitch changes Table IV. However, it is particularly noteworthy that the Conformer model with CTC loss exhibits a more significant degradation in performance relative to its Transducer variant.

At reduced time stretch, there's a universal decline in the performance across all models Table V. However, the Whisper models, particularly the smaller variants, exhibit an extreme version of this drop, prone to generating repetitive phrases in their outputs. This tendency, known as "hallucination," is widely observed in sequence generation models, impacting both ASR [34] and broader language generation [35], leading to a significantly inflated WER. Interestingly, even at a stretch rate of 0.7 — where humans find the audio completely intelligible [36] — there's still a noticeable decline in recognition accuracy for these models.

V. Conclusion

This study's comparison of Whisper, QuartzNet, and Conformer ASR models against acoustic disturbances such as white noise, reverberation, time stretch, and pitch shift in the context of Italian illustrates each model's unique response to different audio challenges. The results highlight the strengths and weaknesses of each model, offering valuable insights for future improvements in ASR robustness. By focusing on Italian, this research provides important insights for optimizing speech recognition technologies across a broader range of languages, guiding efforts toward creating more universally effective and reliable ASR systems that can maintain high accuracy across different linguistic and acoustic environments.

Future work can consider expanding this analysis to a wider range of languages to assess whether the findings hold across different linguistic contexts, as well as investigating advanced noise reduction techniques and developing methodologies to create more robust ASR models. These efforts will further enhance the resilience and applicability of ASR systems across diverse linguistic and acoustic environments.

References

[1] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proceedings of the 34th International Conference on Neural Information Processing Systems*, ser. NIPS'20. Red Hook, NY, USA: Curran Associates Inc., 2020.

[2] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022.

[3] A. Gulati, C.-C. Chiu, J. Qin, J. Yu, N. Parmar, R. Pang, S. Wang, W. Han, Y. Wu, Y. Zhang, and Z. Zhang, Eds., *Conformer: Convolution-augmented Transformer for Speech Recognition*, 2020.

[4] J. Li, V. Lavrukhin, B. Ginsburg, R. Leary, O. Kuchaiev, J. Cohen, H. Nguyen, and R. Gadde, "Jasper: An end-to-end convolutional neural acoustic model," 09 2019, pp. 71–75.

[5] S. Kriman, S. Beliaev, B. Ginsburg, J. Huang, O. Kuchaiev, V. Lavrukhin, R. Leary, J. Li, and Y. Zhang, "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)*, 2020, pp. 6124–6128.

[6] J. Li, L. Deng, Y. Gong, and R. Häb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, pp. 745–777, 2014. [Online]. Available: https://api.semanticscholar.org/CorpusID:14557362

[7] Y. Higuchi, N. Tawara, A. Ogawa, T. Iwata, T. Kobayashi, and T. Ogawa, "Noise-robust attention learning for end-to-end speech recognition," in *2020 28th European Signal Processing Conference (EUSIPCO)*, 2021, pp. 311–315.

[8] T. Cui, J. Xiao, L. Li, X. Jiang, and Q. Liu, "An approach to improve robustness of nlp systems against asr errors," *ArXiv*, vol. abs/2103.13610, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:232352551

[9] J. Huang, O. Kuchaiev, P. O'Neill, V. Lavrukhin, J. Li, A. Flores, G. Kucsko, and B. Ginsburg, "Cross-language transfer learning, continuous learning, and domain adaptation for end-to-end automatic speech recognition," 2020.

[10] M. Mauch and S. Ewert, "The audio degradation toolbox and its application to robustness evaluation," in *International Society for Music Information Retrieval Conference*, 2013. [Online]. Available: https://api.semanticscholar.org/CorpusID:11675708

TABLE II
WER FOR DIFFERENT WHITE NOISE LEVELS

Model / Noise Level	0	0.001	0.002	0.005	0.01	0.015	0.02	0.025	0.03
Whisper Base	0.37	0.41	0.41	0.53	0.70	0.81	0.98	1.10	1.24
Whisper Medium	0.10	0.11	0.12	0.15	0.20	0.24	0.27	0.31	0.36
Whisper Large-v3	0.06	0.06	0.07	0.09	0.11	0.13	0.16	0.19	0.21
Conformer CTC Large	0.07	0.08	0.09	0.11	0.14	0.18	0.21	0.24	0.27
Conformer Transducer Large	0.05	0.06	0.07	0.09	0.11	0.13	0.15	0.17	0.19
FastConformer Hybrid Large	0.06	0.06	0.07	0.09	0.13	0.16	0.20	0.23	0.26
QuartzNet	0.17	0.18	0.19	0.22	0.27	0.31	0.34	0.38	0.41

TABLE III
WER FOR DIFFERENT REVERBERATION TIMES

Model / Reverberation Time (s)	Original	2.0	1.5	1.0
Whisper Base	0.37	1.86	2.21	2.21
Whisper Medium	0.10	0.41	0.40	0.41
Whisper Large-v3	0.06	0.22	0.20	0.20
Conformer CTC Large	0.07	0.59	0.60	0.57
Conformer Transducer Large	0.05	0.52	0.52	0.50
FastConformer Hybrid	0.06	0.55	0.56	0.52
QuartzNet	0.17	0.70	0.72	0.71

TABLE IV
WER FOR PITCH SHIFT TRANSFORMATION

Model / Num steps	Original	1	2	3
Whisper Base	0.37	1.37	1.45	1.55
Whisper Medium	0.10	0.37	0.36	0.39
Whisper Large-v3	0.06	0.21	0.20	0.20
Conformer CTC Large	0.07	0.34	0.36	0.39
Conformer Transducer Large	0.05	0.20	0.20	0.22
FastConformer Hybrid	0.06	0.28	0.30	0.32
QuartzNet	0.17	0.69	0.72	0.75

TABLE V
WER FOR TIME STRETCH TRANSFORMATION

Model / Stretch Factor	Original	0.7x	0.3x	0.1x
Whisper Base	0.37	1.22	3.34	13.38
Whisper Medium	0.10	0.33	0.68	5.22
Whisper Large-v3	0.06	0.18	0.44	1.91
Conformer CTC Large	0.07	0.39	0.98	1.00
Conformer Transducer Large	0.05	0.22	0.88	1.00
FastConformer Hybrid	0.06	0.30	0.97	1.00
QuartzNet	0.17	0.78	0.99	1.00

[11] P. Eickhoff, M. Möller, T. Pekarek-Rosin, J. Twiefel, and S. Wermter, "Bring the noise: Introducing noise robustness to pretrained automatic speech recognition," in *International Conference on Artificial Neural Networks*, 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:261559431

[12] G. W. Lee and H. K. Kim, "Two-step joint optimization with auxiliary loss function for noise-robust speech recognition," *Sensors (Basel, Switzerland)*, vol. 22, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:250942334

[13] J. C. Duarte and S. Colcher, "Building a noisy audio dataset to evaluate machine learning approaches for automatic speech recognition systems," *ArXiv*, vol. abs/2110.01425, 2018. [Online]. Available: https://api.semanticscholar.org/CorpusID:238259030

[14] F. Adolfi, J. S. Bowers, and D. Poeppel, "Successes and critical failures of neural networks in capturing human-like speech recognition," *Neural Netw.*, vol. 162, no. C, p. 199–211, may 2023. [Online]. Available: https://doi.org/10.1016/j.neunet.2023.02.032

[15] U. Shrawankar and V. Thakare, "Noise estimation and noise removal techniques for speech recognition in adverse environment," in *Intelligent Information Processing V*, Z. Shi, S. Vadera, A. Aamodt, and D. Leake, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 336–342.

[16] K. Saito, N. Murata, T. Uesaka, C.-H. Lai, Y. Takida, T. Fukui, and Y. Mitsufuji, "Unsupervised vocal dereverberation with diffusion-based generative models," 2022.

[17] B. Schwartz, S. Gannot, and E. Habets, "Online speech dereverberation using kalman filter and em algorithm," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 23, pp. 394–406, 2015. [Online]. Available: https://api.semanticscholar.org/CorpusID:2413399

[18] J. Balam, J. Huang, V. Lavrukhin, S. Deng, S. Majumdar, and B. Ginsburg, "Improving noise robustness of an end-to-end neural model for automatic speech recognition," 2020.

[19] D. Fucci, M. Gaido, M. Negri, M. Cettolo, and L. Bentivogli, "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. [Online]. Available: https://api.semanticscholar.org/CorpusID:263830339

[20] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, and A. M. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing:*

System Demonstrations. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45. [Online]. Available: https://www.aclweb.org/anthology/2020.emnlp-demos.6

[21] O. Kuchaiev, J. Li, H. Nguyen, O. Hrinchuk, R. Leary, B. Ginsburg, S. Kriman, S. Beliaev, V. Lavrukhin, J. Cook, P. Castonguay, M. Popova, J. Huang, and J. M. Cohen, "Nemo: a toolkit for building ai applications using neural modules," 2019.

[22] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An asr corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[23] C. Cieri, D. Miller, and K. Walker, "The fisher corpus: A resource for the next generations of speech-to-text," 01 2004.

[24] J. Godfrey, E. Holliman, and J. McDaniel, "Switchboard: telephone speech corpus for research and development," in *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, 1992, pp. 517–520 vol.1.

[25] D. B. Paul and J. M. Baker, "The design for the wall street journal-based csr corpus," in *Proceedings of the Workshop on Speech and Natural Language*, ser. HLT '91. USA: Association for Computational Linguistics, 1992, p. 357–362. [Online]. Available: https://doi.org/10.3115/1075527.1075614

[26] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *International Conference on Language Resources and Evaluation*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:209376338

[27] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural 'networks," vol. 2006, 01 2006, pp. 369–376.

[28] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, and R. Collobert, "Mls: A large-scale multilingual dataset for speech research," in *Interspeech 2020*. ISCA, Oct. 2020. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2020-2826

[29] C. Wang, M. Rivière, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. M. Pino, and E. Dupoux, "Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," *CoRR*, vol. abs/2101.00390, 2021. [Online]. Available: https://arxiv.org/abs/2101.00390

[30] Google, "Sentencepiece," https://github.com/google/sentencepiece.

[31] Q. Zhang, H. Lu, H. Sak, A. Tripathi, E. McDermott, S. Koo, and S. Kumar, "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)*, 2020, pp. 7829–7833.

[32] D. Rekesh, N. R. Koluguri, S. Kriman, S. Majumdar, V. Noroozi, H. Huang, O. Hrinchuk, K. Puvvada, A. Kumar, J. Balam, and B. Ginsburg, "Fast conformer with linearly scalable attention for efficient speech recognition," 2023.

[33] K. L. Payton, R. M. Uchanski, and L. D. Braida, "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*, vol. 95, no. 3, pp. 1581–1592, 03 1994. [Online]. Available: https://doi.org/10.1121/1.408545

[34] R. Frieske and B. E. Shi, "Hallucinations in neural automatic speech recognition: Identifying errors and hallucinatory models," 2024.

[35] A. Holtzman, J. Buys, M. Forbes, and Y. Choi, "The curious case of neural text degeneration," *CoRR*, vol. abs/1904.09751, 2019. [Online]. Available: http://arxiv.org/abs/1904.09751

[36] J. A. Müller, D. Wendt, B. Kollmeier, S. Debener, and T. Brand, "Effect of speech rate on neural tracking of speech," *Frontiers in Psychology*, vol. 10, 2019. [Online]. Available: https://www.frontiersin.org/journals/psychology/articles/10.3389/fpsyg.2019.00449