Is Noise Reduction Improving Open-Source ASR Transcription Engines Quality?

Asma Trabelsi, Laurent Werey, Sébastien Warichet and Emmanuel Helbert
Alcatel-Lucent Enterprise, ALE International, 32, avenue Kléber 92700 Colombes, Paris, France

Keywords: Speech Recognition, ASR, Data Sovereignty, Vosk, Whisper.

Abstract: Transcription has becoming an important task on the field of Artificial Intelligence and Machine Learning. Much research has focused on such a field so that we find a lot of paid and open-source ASR solutions. The choose of the best solution is crucial. Open source ones seems to be appropriate especially for companies that would maintain the aspect of data sovereignty. Vosk and Whisper are ASR open-source tools that have been revolutionized this last period. The first idea of this paper is to compare these two solutions in term of Word Error Rate (WER) to conclude who performs best. In the meantime, a lot of models aroused focusing on removing disturbing noises (such as dog barks, child screams, etc) during remote communication. The second idea of the paper is to study the influence of such models applied prior to the transcription service on the quality of the communication transcription. In our study, we focused on voice mail transcription use case.

1 INTRODUCTION

In the sector of communications solutions, NLP service are changing the game, bringing new features and improving performance far beyond hardware or digital communication signal processing. Cloud Communication platforms offer several features to cite a few video or audio call, screen or file sharing, which can be stored on the Cloud. It allows fluid and modern communication inside companies as well as with the external environment. Transcription functionality is a key function brought by NLP. It makes the content of a conversation accessible to everyone including those with hearing loss and allows for additional processing such as language translation, project and data knowledge management (Bain et al., 2005). Today, there are several Automatic Speech Recognition (ASR) solutions. There are paid solutions such as Google, IBM, Microsoft, Linagora (Rebai et al., 2020), etc. There exist also open-source solutions like Mozilla DeepSpeech(Nacimiento-García et al., 2021), Kaldi (Povey et al., 2011), Vosk (alphacephei, 2023), wav2letter (Collobert et al., 2016), SpeechBrain (Ravanelli et al., 2021), Whisper (Radford et al., 2023), etc. Choosing the best solution is a crucial and essential step for companies and will depend on the context of use and operation. In a previous work (Trabelsi et al., 2022), We defined four important criteria (SCQA) for choosing the most appropriate solution.

- **Data Security.** To guarantee data security and comply with the GDPR, we preferred that the calculation (hosting) be done either on premises or on a well-identified cloud server.
- **Cost.** The choice of the solution will also be based on the cost of the operation. Paid solutions charge either per transaction or per model.
- **Quality.** The performance of the models is a significant criterion for ensuring a better service.
- **Adaptability.** The chosen solution must be adjustable to the vocabularies of our customers. The solution must be adapted according to the needs of our customers and according to the business sectors

The majority of paid solutions shows poor results regarding criteria of cost and data sovereignty. There are non global suppliers ensuring better data security and opening up to more flexible business models but their Quality criterion is poor as they do not support many languages. For an international company, this criterion is critical. Hence, the best trad-off seemed to use open-source models such as Vosk (alphacephei, 2023) and SpeechBrain (Ravanelli et al., 2021) that have revolutionized in recent times. In (Trabelsi et al., 2022), we conducted an experimental study to compare these two solutions based on the Word Error Rate

Trabelsi, A., Werey, L., Warichet, S. and Helbert, E.
Is Noise Reduction Improving Open-Source ASR Transcription Engines Quality?.
DOI: 10.5220/0012457100003636

In *Proceedings of the 16th International Conference on Agents and Artificial Intelligence (ICAART 2024) - Volume 3*, pages 1221-1228
ISBN: 978-989-758-680-4; ISSN: 2184-433X

(WER) and the Inference Time (IT). We have experimentally proven that Vosk outperforms SpeechBrain in both WER and IT critertia. At the beginning of 2023, we experienced the revolution following OpenAI tools. Among its tools, we identify additional open-source solutions for transcription with the release of Whisper. It became then highly relevant to compare the results delivered by Whisper with those of our first champion, Vosk.
In parallel to this work, we did some experimentation of ML models to improve the overall quality of communication by adding features to reduce or event remove external noises during conversation. That approach was definitely in the scope of the Q criterion of our framework and it raised the question of the impact of this pre-processing on voice transcription. Will the noise reduction decrease the performance of the transcription or on the contrary, will it improve the WER of the transcription model? Our hypothesis is that a noise reduction treatment on the audio content can increase the transcription quality. Thus, the second aim of this paper is to study the impact of noise reduction on transcription quality based on the Word Error Rate. We consider two well-known noise reduction frameworks which are RNNoise(Mozzila, 2023) and ASTEROID (Asteroid, 2023) that will be applied to the winner ASR framework among Vosk and Whisper.

The remaining of this paper is organized as follows. In section 2, we focus on existing open source solutions by highlighting their advantages and their drawbacks. In Section 3, we present some well-known noise reduction techniques. We explain the experimentation settings and results in Section 4. Section 5 discussed the results obtained and we draw the conclusion and our perspectives in Section 6.

2 ASR OPEN SOURCE FRAMEWORK

Open-source tools seem particularly well suited to meet our SCQA criteria. Kaldi (Povey et al., 2011), Vosk (alphacephei, 2023) and Whisper (Radford et al., 2023) have been remarkably successful in both academical and industrial areas. We detail below each of those frameworks.

2.1 Kaldi and Vosk

Kaldi (Povey et al., 2011) is one among the well known framework that has seen a lot of success recently. Several ASR companies have been developing their offer using this framework such as Linagora (Rebai et al., 2020) to cite a few. It is a comprehensive toolkit for speech recognition that has gained popularity in both research and industry. It provides a wide range of tools and resources for building and customizing ASR systems. Kaldi is known for its flexibility and extensibility, allowing researchers and developers to experiment with various ASR components, such as acoustic modeling, language modeling, and decoding algorithms. It is designed with a modular architecture, making it easier to experiment with different ASR components and integrate external libraries. It includes state-of-the-art models and training techniques for acoustic and language modeling. It supports multiple languages and it can be tailored to specific languages and dialects. Vosk (alphacephei, 2023) is a well known open-source solution for ASR that is build on top of Kaldi. It is a lightweight and efficient ASR tool that is built with a focus on speed, accuracy, and ease of use. It comes with pre-trained models for multiple languages, simplifying the setup process. Vosk is designed to be resource-efficient, making it a good choice for resource constrained devices and it is particularly well-suited for applications where low-latency speech recognition is required, such as online transcription, voice assistants, etc (Gentile et al., 2023).

2.2 Whisper and Whisper-Faster

At the beginning of 2023, we experienced a very remarkable evolution with the arrival of OpenAI solutions, notably ChatGPT and its newest Large Language Models (LLM) (Mao et al., 2023). Additionally, OpenAI has succeeded in improving its transcription engine called Whisper (Radford et al., 2023). It represents a significant advancement in the field of Automatic Speech Recognition. Whisper's capabilities have been harnessed in a variety of applications, from transcription services to voice assistants (Mul, 2023; Spiller et al., 2023). One of the key strengths of Whisper lies in its ability to adapt and perform exceptionally well across multiple languages and accents, making it a versatile tool for speech-to-text conversion on a global scale. Its training data, which comprises a vast and diverse dataset of multilingual and multitask supervised data, allows it to continually improve its accuracy and robustness. Researchers and developers have also been exploring ways to fine-tune Whisper for specific applications, such as medical transcription and customer service call analytics (Jain et al., 2023). This adaptability and scalability make Whisper a valuable asset in fields where accurate speech recognition is crucial, revolutionizing the way we interact with spo-

ken language data. For online ASR service, real-time ASR systems like Whisper, improving the speed at which speech is recognized and transcribed is an ongoing challenge. Faster-Whisper is a reimplementaion of OpenAI's Whisper model, built upon the robust CTranslate2 framework, a high-speed inference engine tailored for Transformer models (Macháček et al., 2023). Compared to the original OpenAI Whisper, this implementation delivers a remarkable speed boost, achieving up to fourfold faster performance without compromising accuracy. Furthermore, it accomplishes this feature while consuming fewer system resources, making it a more efficient choice. Additionally, the implementation offers room for optimization through the application of 8-bit quantization on both CPU and GPU, further enhancing its efficiency (Fas,). Researchers are constantly working on optimizing the processing time of ASR systems to ensure faster real-time speech-to-text conversion. As Whipser has not been designed in first place for real time transcription, some open source solutions, built on top of Faster-Whispser, exist today like Whisper streaming (Macháček et al., 2023).

3 NOISE REDUCTION TECHNIQUES

The world of noise cancellation and audio signal processing is vast and constantly evolving (Benesty et al., 2009). Each year, conferences such as ICASSP see the publication of dozens of research papers dealing with the subject, testifying to its growing importance. Apart from scientific conferences, annual challenges like the Deep Noise Suppression Challenge inspire the scientific community to innovate and propose cutting-edge solutions for noise suppression. Major players such as Microsoft, Amazon and Baidu play a key role, often introducing new approaches and methodologies. Noise cancellation is a crucial technology in audio, operating in both real time or offline audio signals. It aims to filter and eliminate all impairing sounds that could be either background noise, related to the audio environment, or communication noise, related to the communication infrastructure. This greatly facilitates communications distance, guaranteeing better listening quality and reducing inconvenience for users. These noise cancellation techniques could be either based on Digital Signal Processing (Vaseghi, 2008) or on Machine Learning (ML) algorithms. In this paper, we focus only on ML algorithms. One of the major problems introduced by this mechanism is latency, especially when it comes to deal with real time audio signals. Latency is the delay time between capturing an audio segment and returning it once processed. Excessive latency can significantly affect the quality of a conference dialogue, increasing the overall round trip delay and making the conversation less fluid and natural (Suznjevic and Saldana, 2016). It will therefore be necessary to ensure that latency is as low as possible on a set of devices with uncontrolled computing powers. In this section, we will present two open-source solutions: RNNoise(Mozzila, 2023) and ASTEROID (Asteroid, 2023). RNNoise is designed to cope with both real time and offline audio signal and to guarantee low latency in case of real time noise reduction. ASTEROID only treats noise reduction in offline mode.

3.1 RNNoise

RNNoise is a valuable tool for enhancing audio quality by effectively reducing unwanted noise and disturbances in audio recordings (Mozzila, 2023; Valin, 2018). It is an open-source project available under BSD-3-Clause licence meaning that its source code is freely available to the public even for commercial use (RNNoise, 2023). This encourages collaboration, improvement, and integration into various software applications. RNNoise employs a machine learning-based approach to distinguish between speech and noise in audio signals. It uses neural networks to classify and separate these components, which helps in effectively reducing background noise. It is mainly designed to work in real-time, making it suitable for applications like voice and video calls, online conferencing, and live audio streaming where immediate noise reduction is required. It is optimized for low-latency performance, which is crucial in applications like real-time voice communication, where delays can negatively impact the user experience. Another advantage is that RNNoise can be integrated into various software and hardware systems, making it versatile and adaptable to different use cases. It is commonly used in voice communication software, speech recognition systems, and audio post-processing tools. Users can often adjust parameters and settings within RNNoise to fine-tune the noise reduction process according to their specific requirements and the nature of the audio source. One drawback of RNNoise is its incapacity to work with various audio formats. It operates only on RAW 16-bit (machine endian) mono PCM files sampled at 48 kHz. Another point is that the model operates using 22 filter bands quite wide, meaning it cannot isolate or remove noise with narrow frequency bandwidth located in the voice frequencies band. As a result, some noise remains in the output.

Additionally, the model showed notable sensitivity to variations of input volume and voice characteristics, such as timbre and pitch.

3.2 ASTEROID

ASTEROID is an open-source toolkit (Pariente et al., 2020) designed to facilitate deep learning-based audio source separation and speech enhancement, catering to both researchers and industrials. Built using PyTorch, a highly popular dynamic neural network toolkit, ASTEROID prioritizes user-friendliness, extensibility, promoting reproducible research, and fostering effortless experimentation. Consequently, it accommodates a diverse array of datasets and architectures and includes pre-configured setups to replicate significant research papers. ASTEROID is publicly accessible on (Asteroid, 2023). ASTEROID propose audio source separation models such as Deep clustering (Hershey et al., 2016), ConvTasNet (Luo and Mesgarani, 2019), DPRNN (Luo et al., 2020) and others (HuggingFace, 2023). In addition to the models themselves, ASTEROID offers essential components like building blocks, loss functions, metrics, and frequently employed datasets in the field of source separation. This simplifies the process of creating novel source separation models and facilitates their comparative evaluation against existing ones.

4 EXPERIMENTAL COMPARISON AND RESULTS

4.1 Experimentation Settings

Data collection is still the most important task in any machine learning application. GDPR rules for data privacy and data governance imposes to collect carefully our data. We did not find public and open dataset containing relevant business voice messages adapted to our experimentation. Hence, the data collection was done manually by asking volunteers within the company to send us some professional voice messages. This procedure has allowed us to obtain 74 voice messages (47 French voice messages and 27 English voice messages). For the purpose of studying the impact of noise reduction on the transcription quality, we generate denoised voice messages through RNNoise and ASTEROID. For ASTEROID denoised tool, we have used specifically the DCCRNet Libri1Mix Enhsingle 16k model available on HuggingFace. This model was preferred owing to its proven efficacy in the domain of single voice enhancement compared to all other tested models.

Two metrics MOS (Union, 2016) and WER were pivotal in our evaluations. The Mean Opinion Score (MOS) provides an appraisal of the perceived quality of an audio communication, rated on a scale from 1 (very poor quality) to 5 (excellent quality). It usually aggregates the evaluations of multiple individuals, offering a holistic sense of how the audio might be perceived by an average listener. In our case we chose to use DNSMOS (objective speech quality metric to evaluate noise suppressors) to avoid setting the protocol of MOS and compute it objectively. The MOS metric is chosen for its capacity to offer a comprehensive view and to lend insight into the real-world implications and user perception. The WER reflects the Word Error Rate and it will be used to qualify the ASR engine. In our case, we only consider Vosk and Whisper ASR engines.

4.2 Experimentation Results

In what follows, we present our experimentation results. We firstly compare Vosk and Whisper in term of WER for both French and English data. Then, we study the impact of noise reduction on transcription quality for Whisper models.

4.2.1 Vosk versus Whisper in Term of WER

Figure 1 and Figure 2 presents the WER for various models of Whisper and Vosk and for two languages, French and English. While Whisper is able to cope with various language with the same model, we had to configure Vosk to use specific language models. We tested one single model in French and four different models in English. On the Whisper side, we varied the size of the model. If we compare the results between the French and English languages, Vosk with the French model performs better than for English models. At the opposite, Whisper performs better in English language. Now if we only consider the different Whisper models, as expected and without surprise, the experimentation confirmed that the larger the model, the better the results. With the drawback that large models will require more computation resources and lead to higher latency. Finally, when comparing Vosk and Whisper figures, there is clear and impressive better performances for all Whisper models in English with WER below 25% even for whisper-small.

4.2.2 MOS and Whisper WER for Original and Denoised Audio

- **MOS Results.**
  Before presenting the Whisper WER for origi-

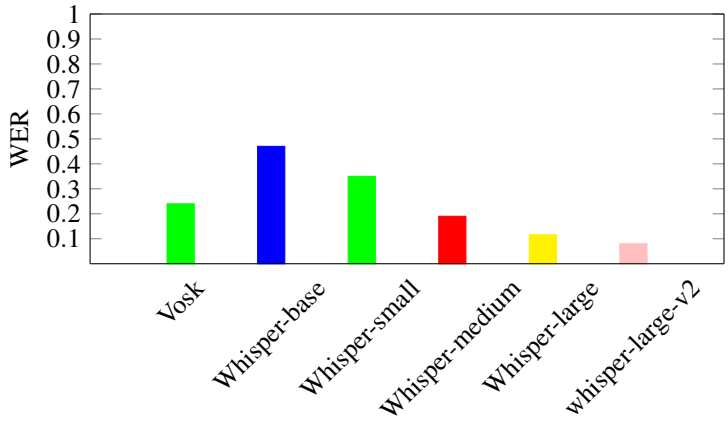

Figure 1: Vosk Vs Whisper for French data in term of WER.

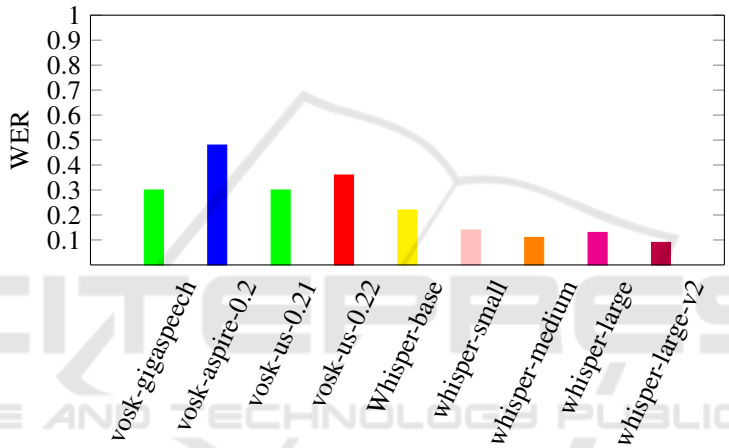

Figure 2: Vosk Vs Whisper for English data in term of WER.

nal and denoised audio, we present the MOS to assess the impact of the noise suppression models. Figure 3 and Figure 4 depict the results for both French en English voice messages. Though we cannot extract a clear prediction of the impact of the denoising, there are some trends which stem from these preliminary results. ASTEROID perfoms better than RNNoise for voice messages with MOS above 3,1 (approximated threshold). While ASTEROID improve the overall MOS score for these samples, RNNoise degrade the quality, decreasing significantly the MOS. But for samples with poor quality (MOS lower than 3,1), RNNoise performs definitely better than ASTEROID, improving the overall score. This behaviour is independent from the language.

- **WER Results.**

  Figure 5 presents the WER results for original and denoised audio using the different whisper model from base to large-v2. We decided to consider all the voice messages for the WER calculation and the figures correspond to the mean WER. This approach allow us to sense the general trend of the impact of denoising on transcription quality, whatever the initial MOS score. If we compare the impact of denoising depending on the Whisper model size, there is a real improvement of the transcription quality for base and small models while, at the opposite, ASTEROID and RNNoise degrade the overall transcription quality for larger models. For Whisper base and Whisper small, ASTEROID performs slightly better than RNNoise but one must recall that RNNoise tends to degrade the MOS score for message with already good audio quality. For Whisper medium, Whisper large and Whisper large-v2, RNNoise performs better than ASTEROID with only a slight degradation of the WER.

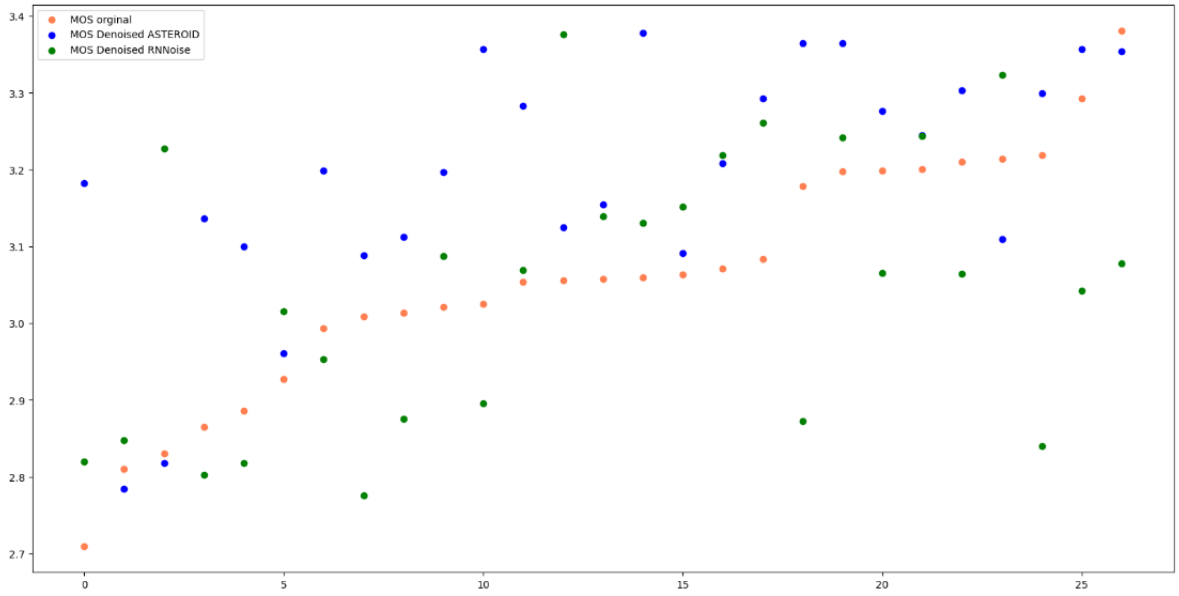

Figure 3: MOS for English dataset.

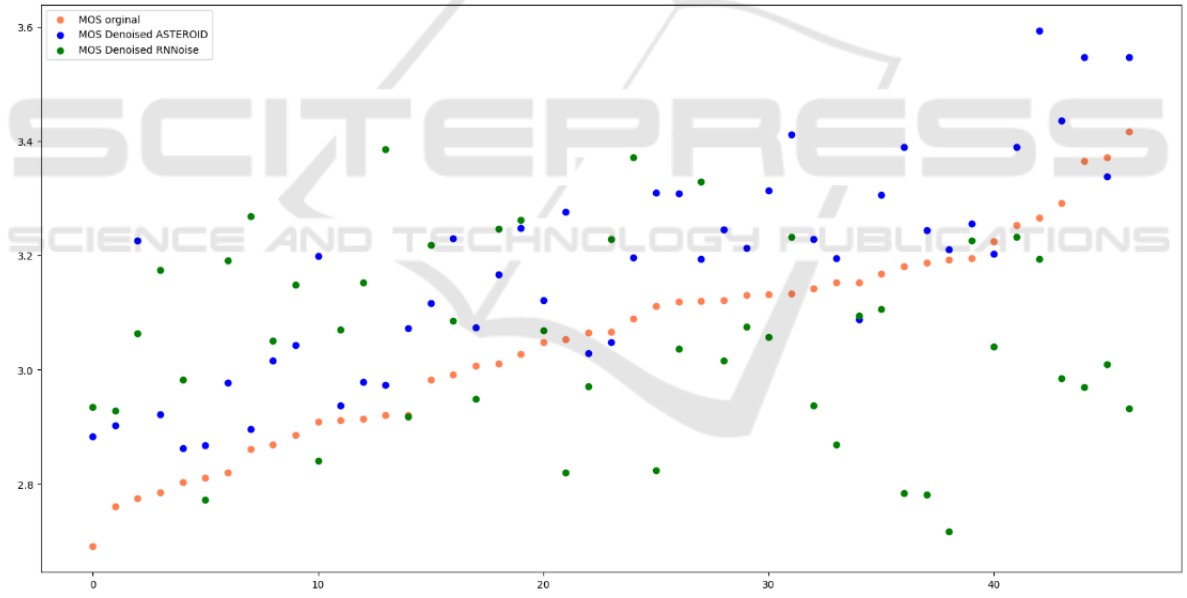

Figure 4: MOS for French dataset.

5 DISCUSSION

We have experimentally proven that Whisper models, even the small ones outperform all Vosk models for English. For the French language, Vosk has outperformed Whisper base and small models but Whisper medium, large and large v2 models have given better results than Vosk. In short, for French and English language, the overall assessment is that Whisper performs better than Vosk in terms of Word Error Rate. We have next studied the impact of noise reduction on the audio quality when based on the MOS criterion. Two noise reduction approaches have been used, RNNoise and ASTEROID. We have shown that ASTEROID improves audio quality whatever the initial MOS score while RNNoise behaviour is a little less deterministic. While RNNoise outperforms ASTEROID for audio messages with poor quality (MOS

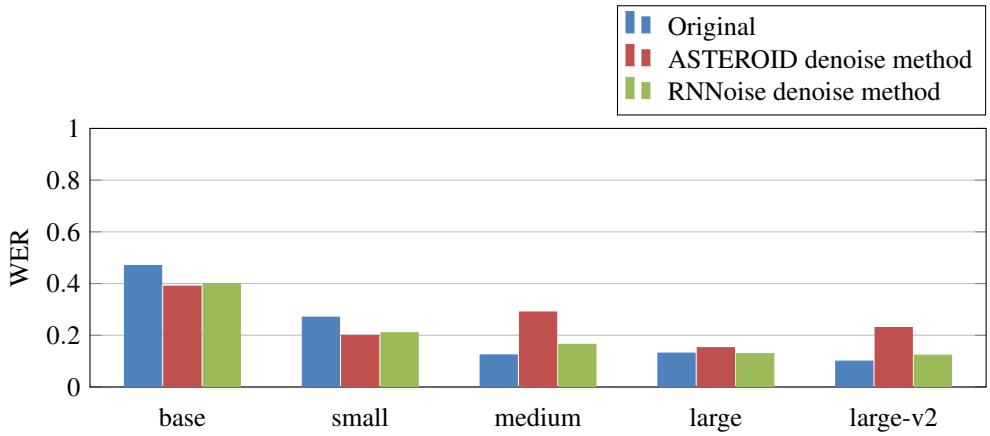

Figure 5: WER for noised and denoised audio.

lower than 3.1), for audio messages with relative good quality, RNNoise behaves in the opposite way to what is expected. This is may be due to the fact that RNNoise is more adequate for live streaming with sample rate equals to 48khz while the voice mail audio collected have 16khz as sample rate. When using RNNoise, we should upsample our audio to 48khz to reduce background noise and then we need to downsample to 16khz in order to compute the MOS. Once noise reduction is done, we have applied Whisper to get transcription. Results have proven that the WER for base and small models of denoised files with both ASTEROID and RNNoise is lower than the WER obtained with the original audio. For base and small models, the WER obtained for denoised files with ASTEROID is a little bit less than for the denoised files with RNNoise. For medium, large and large-v2 models, original files have given the least WER and files denoised with RNNoise have given a lower WER than files denoised with ASTEROID. At this stage, we can not conclude that improving the overall MOS score of voice audio has a positive impact on the transcription quality.

6 CONCLUSION

In this paper, we have compared Vosk and Whisper using the WER criteria. Experimentation results have proven the better performance of Whisper over Vosk. We have then put the focus on the impact of noise reduction on the transcription quality. Results proved that noise reduction can have a positive impact, especially for small language models. This may be explained by the fact that these models are trained with a small amount of data that does not take into consideration noisy data contrary to the larger models (Medium, Large and Large-V2 for Whisper transcription engine). For future work, we would like to augment the dataset by collecting more voice mail data with various MOS and various type of audio impairments. The purpose is to understand more in detail what kinds of audio improvement brought by noise reduction frameworks have the best impact on the transcription WER. Finally, We also would like to compare models and study the impact of noise reduction for languages other than French and English.

REFERENCES

Faster-whisper. https://github.com/guillaumekln/faster-whisper.

alphacephei (2023). Vosk. https://alphacephei.com/nsh/.

Asteroid (2023). Github Asteroid. https://github.com/asteroid-team/asteroid.

Bain, K., Basson, S., Faisman, A., and Kanevsky, D. (2005). Accessibility, transcription, and access everywhere. *IBM systems journal*, 44(3):589–603.

Benesty, J., Chen, J., Huang, Y., and Cohen, I. (2009). *Noise reduction in speech processing*, volume 2. Springer Science & Business Media.

Collobert, R., Puhrsch, C., and Synnaeve, G. (2016). Wav2letter: an end-to-end convnet-based speech recognition system. *arXiv preprint arXiv:1609.03193*.

Gentile, A. F., Macrì, D., Greco, E., and Forestiero, A. (2023). Privacy-oriented architecture for building automatic voice interaction systems in smart environments in disaster recovery scenarios. In *2023 International Conference on Information and Communication Technologies for Disaster Management (ICT-DM)*, pages 1–8. IEEE.

Hershey, J. R., Chen, Z., Le Roux, J., and Watanabe, S. (2016). Deep clustering: Discriminative embeddings for segmentation and separation. In *2016 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 31–35. IEEE.

HuggingFace (2023). Hugging Face Asteroid. https://huggingface.co/models?library=asteroid.

Jain, R., Barcovschi, A., Yiwere, M., Corcoran, P., and Cucu, H. (2023). Adaptation of whisper models to child speech recognition. *arXiv preprint arXiv:2307.13008*.

Luo, Y., Chen, Z., and Yoshioka, T. (2020). Dual-path rnn: efficient long sequence modeling for time-domain single-channel speech separation. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 46–50. IEEE.

Luo, Y. and Mesgarani, N. (2019). Conv-tasnet: Surpassing ideal time–frequency magnitude masking for speech separation. *IEEE/ACM transactions on audio, speech, and language processing*, 27(8):1256–1266.

Macháček, D., Dabre, R., and Bojar, O. (2023). Turning whisper into real-time transcription system. *arXiv preprint arXiv:2307.14743*.

Mao, R., Chen, G., Zhang, X., Guerin, F., and Cambria, E. (2023). Gpteval: A survey on assessments of chatgpt and gpt-4. *arXiv preprint arXiv:2308.12488*.

Mozzila (2023). RNNoise: Learning Noise Suppression. https://jmvalin.ca/demo/rnnoise/.

Mul, A. (2023). Enhancing dutch audio transcription through integration of speaker diarization into the automatic speech recognition model whisper. Master's thesis.

Nacimiento-García, E., González-González, C. S., and Gutiérrez-Vela, F. L. (2021). Automatic captions on video calls, a must for the elderly: Using mozilla deepspeech for the stt. In *Proceedings of the XXI International Conference on Human Computer Interaction*, pages 1–7.

Pariente, M., Cornell, S., Cosentino, J., Sivasankaran, S., Tzinis, E., Heitkaemper, J., Olvera, M., Stöter, F.-R., Hu, M., Martín-Doñas, J. M., et al. (2020). Asteroid: the pytorch-based audio source separation toolkit for researchers. *arXiv preprint arXiv:2005.04132*.

Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlicek, P., Qian, Y., Schwarz, P., et al. (2011). The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, number CONF. IEEE Signal Processing Society.

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Ravanelli, M., Parcollet, T., Plantinga, P., Rouhe, A., Cornell, S., Lugosch, L., Subakan, C., Dawalatabad, N., Heba, A., Zhong, J., et al. (2021). Speechbrain: A general-purpose speech toolkit. *arXiv preprint arXiv:2106.04624*.

Rebai, I., Benhamiche, S., Thompson, K., Sellami, Z., Laine, D., and Lorré, J.-P. (2020). Linto platform: A smart open voice assistant for business environments. In *Proceedings of the 1st International Workshop on Language Technology Platforms*, pages 89–95.

RNNoise (2023). Github RNNoise. https://github.com/xiph/rnnoise.

Spiller, T. R., Ben-Zion, Z., Korem, N., Harpaz-Rotem, I., and Duek, O. (2023). Efficient and accurate transcription in mental health research-a tutorial on using whisper ai for sound file transcription.

Suznjevic, M. and Saldana, J. (2016). Delay limits for real-time services. *IETF draft*.

Trabelsi, A., Warichet, S., Aajaoun, Y., and Soussilane, S. (2022). Evaluation of the efficiency of state-of-the-art speech recognition engines. *Procedia Computer Science*, 207:2242–2252.

Union, I. T. (2016). Mean opinion score interpretation and reporting. Standard, International Telecommunication Union, Geneva, CH.

Valin, J.-M. (2018). A hybrid dsp/deep learning approach to real-time full-band speech enhancement. In *2018 IEEE 20th international workshop on multimedia signal processing (MMSP)*, pages 1–5. IEEE.

Vaseghi, S. V. (2008). *Advanced digital signal processing and noise reduction*. John Wiley & Sons.