Improving audio of emergency calls in Spanish performed to the ECU 911 through filters for ASR technology

Marcos Orellana[1 [0000−0002−3671−9362]], Angel Alberto Jimenez Sarango[1 [0000−0003−0018−4535]] and Jorge Luis Zambrano-Martinez[1 [0000−0002−5339−7860]]

Universidad del Azuay, Cuenca, Ecuador
{marore, jorge.zambranno}@uazuay.edu.ec, ajimenez@es.uazuay.edu.ec
https://lidi.uazuay.edu.ec

Abstract. In recent years, Automatic Speech Recognition (ASR) services have performed notable progress in the research efforts of big companies such as Google and Amazon. However, the ASRs are still sensitive to the audio processing quality in other languages. To solve this issue, various speech enhancement algorithms that are the most prominent in improving speech intelligibility were proposed, such as Singular Value Decomposition (SVD), log Minimum Mean Square Error (log-MMSE) and Wiener. By preprocessing the audio files with these algorithms, we seek to reduce the Word Error Rate (WER), which compares the transcription performed by the ASR against a manual transcription. Thus, we can determine the percentage of error that the ASR service has acquired. Results demonstrated that Google is more efficient than Amazon and Vosk counterparts. Also, we decided that applying a Low-pass filter combined with a log-MMSE algorithm to the audio files can substantially reduce the WER percentage of transcription depending on the noise characteristics contained in the audio.

Keywords: Automatic Speech Recognition · Word Error Rate · speech enhancement algorithms · audio quality improvement.

1 Introduction

Automatic Speech Recognition (ASR) has advanced rapidly in the last few years due to continuous improvements. These systems' quality is affected by features used to record the audios and how they are processed by these algorithms, thus reducing their efficiency in speech recognition [14]. For the correct interpretation of the content of audio files in the ASR algorithms, it is necessary to improve the audio quality through processing techniques [17]. Speech enhancement algorithms such as Singular Value Failure (SVD), Wiener or the Minimum Mean Square Error (MMSE) are the most prominent algorithms due to their performance in improving speech intelligibility [9]. Likewise, other techniques seek to improve the audio quality, such as the low-pass, band-pass or high-pass filter,

which enhances the audio quality in real-time while being more efficient than other audio improvement algorithms [16]. There is a wide variety of research about comparing the performance of ASR algorithms by applying the Word Error Rate (WER) index, a measure commonly used to evaluate those algorithms [7],[8], [15]. Also, other researchers have compared the different speech enhancement algorithms [2], [4], [5], but so far, there is no research about ASR services and applying speech enhancement algorithms. Our proposal focuses on preprocessing audio file datasets of emergency calls provided by the Integrated Security Service ECU 911 for using ASR algorithms after the results are compared with transcriptions performed by humans.

2 Related Works

In a speech-to-text conversion, over 73,7% use the WER metric as an evaluation method for the ASR voice recognition [1], and the ASRs are used in some areas such as telephony, military and client services [6]. However, Kepuska & Bohouta [7] demonstrate that Google API is more efficient than open-source APIs such as Microsoft Speech API or Sphinx-4. Authors calculate the WER index by processing English-language audio from different sources. Several studies compare the performance of ASR services as Vascones et al. [15] demonstrate that Amazon Transcribe has a lower average WER percentage of the transcriptions of audio files without any speech enhancement algorithms from the Integrated Security Service ECU 911 than its competitor Google Cloud Speech-to-Text. In another study, Plaza et al. [12] implement a recognition voice model in Spanish with an ASR offline called CMUSphinx.

Nonetheless, Nian et al. [11] conclude that eliminating the noise in an audio file through background noise removal preprocessing helps decrease the WER index by up to 22.1%. In the same way, Shrawankar & Thakare [13] conclude that the traditional Wiener and log Minimum Mean Square Error (log-MMSE) algorithms are frequently used in speech accuracy tests. However, Chen et al. [3] propose a SVD algorithm to remove noise from audio files. Modhave et al. [10] demonstrate that the Wiener algorithm improves speech quality because this algorithm greatly helps to estimate the noise signal in audio. And Meiniar et al. [10] conclude that it is possible to filter the human speech using a band-pass filter and lose very little speech in the audio files analysed. Thus, we propose preprocessing the audio files pro- vided by the Integrated Security Service ECU 911 through the SVD, log-MMSE, Wiener, and low-pass algorithms that have been demonstrated in the research as the best algorithms for improving speech quality and then transcribing them with a low index-WER.

3 Methodology

This section describes the procedure followed to improve the speech quality audio files in four activities, detailed below.

The first step required creating a Google Cloud account to use the Google Speech-to-Text service. Likewise, an Amazon Web Services account must be created to use the Amazon Transcribe service. Both platforms were implemented with a necessary internal configuration, such as creating workspaces for transcriptions and storage spaces called Buckets on the respective media.

For audio preprocessing, an algorithm was created with Python programming language in its version 3.9.9, including all the speech improvement algorithms selected after a rigorous analysis of the previously reviewed scientific papers. This way, the speech enhancement algorithms SVD, Wiener and log-MMSE were selected. On the other hand, we analysed that using a low-pass filter can help reduce the background noise from audio files, and the processing was performed with this filter.

In this methodology step, a script was created with the necessary configurations to implement the chosen transcription services: Google Cloud Speech-to-Text, Amazon Transcribe and Vosk. In addition, cloud storage services called buckets and configurations were used in the cloud service consoles. Moreover, the script for using Vosk's offline service was developed.

One of the metrics to measure the level of transcripts in an ASR system is the WER index. This rate takes a reference transcript that contains no apparent error because it is done manually. Subsequently, the automatic transcriptions performed by Google, Amazon and Vosk were taken as hypotheses. Therefore, the WER index compares both texts considering the words that exist in the reference transcript (N), which had been inserted (I), deleted (D), substituted (S), and presented as a result, besides the percentage error of the transcriptions concerning the reference text. The WER percentage is shown with the following equation: $WER = \left(\frac{I+D+S}{N}\right) \cdot 100$.

4 Experimental Results

As explained in the previous section, the standard metric used to compare the accuracy of the transcripts produced by the ASRs systems is obtained through the WER equation.

In Figure 1, we can observe the results obtained with the WER index of the transcripts of the ASR systems from a set of audio files provided by the Integrated Security Service ECU 911, jointly with the employment of the audio quality improvement filters. These audio files contain sensitive situations, and they are under personal data protection law. In our experiments, Google Cloud performed better using original audio files, reducing the WER percentage. However, the combination of speech enhancement algorithms did not substantially reduce the WER percentage because the audio quality and speech intelligibility were not equivalent. For this reason, the applied algorithms did not distinguish between background noise and the different natural distortions of the voice. Those algorithms tended to eliminate them equally, reducing speech intelligibility in the audio files studied.

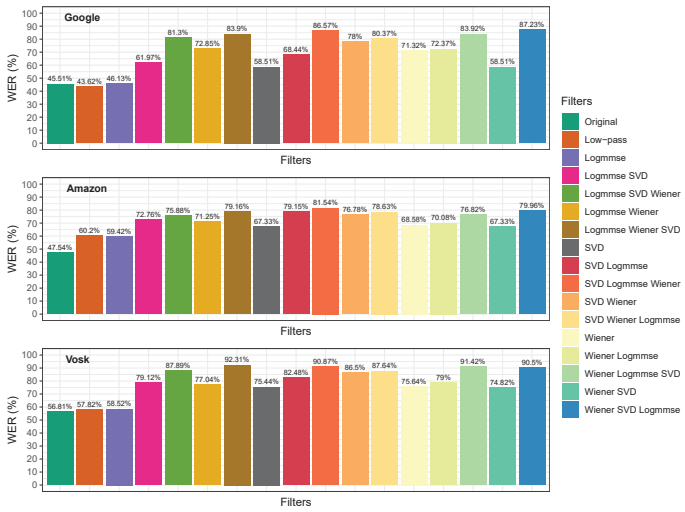

Fig. 1. WER percentage of audio transcripts from ASR systems.

When analysing the results of each audio file transcribed, we observed that a transformed audio file presented a lower WER percentage than the transcription of the original audio file. Other audio files had their WER percentage higher. Figure 2 shows WER percentages of the audio transcripts with the lowest percentages. In comparison, Figure 2 shows the WER of the audio that has the highest rates.

These audio files tested have characteristics that determine the variation of WER percentage. For this reason, audio files with a duration of over three minutes, with a calm conversation between the operator and the caller in an environment with no background noise and with a vocalisation of all the words adequately without the use of a particular lexicon in the Spanish, can identify a significant percentage of words in the audio file. Preprocessing the audio files with the speech enhancement algorithms before applying them to the ASR system as SVD and log-MMSE, they estimated that the noise in the audio signal was not much. The SVD algorithm eliminated part of the speech of an audio file when eliminating the noise from the wave. Although the log-MMSE algorithm estimated signal to noise and the low-pass filter only cleaned the frequencies, the WER percentage was not significantly increased. However, audio files with less than one minute with a conversation without vocalisation adequate, with background noise, and with a lot of regional lexica induced the ASRs to be unable to identify the words optimally and increased the WER percentage. However, log-MMSE and SVD algorithms calculated the excessive noise in the original audio files and removed the noise. Those algorithms did not eliminate the speech contained in the audio file. Applying the low-pass filter removed only frequencies other than the human voice. Hence, the WER index was significantly reduced.

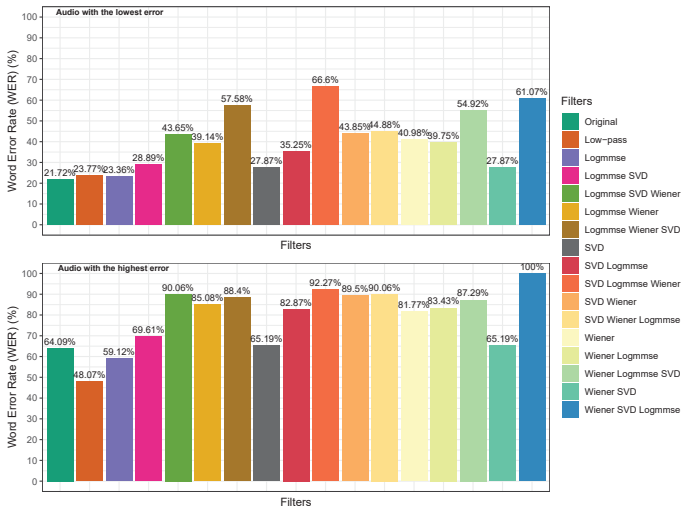

Fig. 2. WER percentage of the audio with the lowest and the highest error.

5 Conclusions

Our research compares transcription quality between the ASR systems offered by Google, Amazon and Vosk after processing the audio files using speech enhancement algorithms such as SVD, log-MMSE, and Wiener and the application of a low-pass filter. The results demonstrate that using ASR offered by Google has a better performance both in the original audio files and in the audios processed by the enhancement algorithms. The WER percentage is reduced depending on the characteristics of the audio files tested, such as the level of background noise, use of a particular lexicon in Spanish, and the length of conversation in the audio files. Using the result obtained as a basis for future work, we plan to develop a classification process for the audio files depending on their characteristics that can be known in advance if one of these filters should be applied, thus considerably improving the transcription of the audio file.

Acknowledgement

This work was partially supported by the vice rectorate of Research at Universidad del Azuay for their financial and academic support, as well as the entire staff in the Computer Science department and the Laboratorio de Investigación y Desarrollo en Informática - LIDI.

References

1. Bhardwaj, V., Ben Othman, M.T., Kukreja, V., Belkhier, Y., Bajaj, M., Goud, B.S., Rehman, A.U., Shafiq, M., Hamam, H.: Automatic speech recognition (asr)

systems for children: A systematic literature review. Applied Sciences **12**(9), 4419 (2022)
2. Chaudhari, A., Dhonde, S.: A review on speech enhancement techniques. In: 2015 International Conference on Pervasive Computing (ICPC). pp. 1–3. IEEE (2015)
3. Chen, X., Litvinov, Y.A., Wang, M., Wang, Q., Zhang, Y.: Denoising scheme based on singular-value decomposition for one-dimensional spectra and its application in precision storage-ring mass spectrometry. Physical Review E **99**(6), 063320 (2019)
4. Dash, T.K., Solanki, S.S.: Comparative study of speech enhancement algorithms and their effect on speech intelligibility. In: 2017 2nd International conference on communication and electronics systems (ICCES). pp. 270–276. IEEE (2017)
5. Gael, P., Chandra, M., Saxena, P., Gupta, V.K.: Comparative analysis of speech enhancement methods. In: 2013 Tenth International Conference on Wireless and Optical Communications Networks (WOCN). pp. 1–5. IEEE (2013)
6. Jamal, N., Shanta, S., Mahmud, F., Sha'abani, M.: Automatic speech recognition (asr) based approach for speech therapy of aphasic patients: A review. In: AIP Conference Proceedings. vol. 1883, p. 020028. AIP Publishing LLC (2017)
7. Këpuska, V., Bohouta, G.: Comparing speech recognition systems (microsoft api, google api and cmu sphinx). Int. J. Eng. Res. Appl **7**(03), 20–24 (2017)
8. Kimura, T., Nose, T., Hirooka, S., Chiba, Y., Ito, A.: Comparison of speech recognition performance between kaldi and google cloud speech api. In: International Conference on Intelligent Information Hiding and Multimedia Signal Processing. pp. 109–115. Springer (2018)
9. Loizou, P.C.: Speech enhancement: theory and practice. CRC press (2007)
10. Modhave, N., Karuna, Y., Tonde, S.: Design of multichannel wiener filter for speech enhancement in hearing aids and noise reduction technique. In: 2016 Online International Conference on Green Engineering and Technologies (IC-GET). pp. 1–4. IEEE (2016)
11. Nian, Z., Tu, Y.H., Du, J., Lee, C.H.: A progressive learning approach to adaptive noise and speech estimation for speech enhancement and noisy speech recognition. In: ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 6913–6917. IEEE (2021)
12. Plaza Salto, J.G., Cristina, S.Z., Acosta Urigüen, M.I., Orellana Cordero, M.P., Cedillo Orellana, I.P., Zambrano-Martinez, J.L.: Speech recognition based on spanish accent acoustic model. Enfoque UTE **13**(3), 45–57 (07 2022)
13. Shrawankar, U., Thakare, V.: Noise estimation and noise removal techniques for speech recognition in adverse environment. In: International Conference on Intelligent Information Processing. pp. 336–342. Springer (2010)
14. Singh, N., Agrawal, A., Khan, R.A.: A critical review on automatic speaker recognition. Science Journal of Circuits, Systems and Signal Processing **4**(2), 14–17 (2015)
15. Vásconez, J.J.P., Ortiz, C.A.N., Cordero, M.P.O., León, P.A.P., Orellana, P.C.: Evaluación del reconocimiento de voz entre los servicios de google y amazon aplicado al sistema integrado de seguridad ecu 911. Revista Tecnológica-ESPOL **33**(2), 147–158 (2021)
16. Wan, F., Yuan, Z., Ravelo, B., Ge, J., Rahajandraibe, W.: Low-pass ngd voice signal sensoring with passive circuit. IEEE Sensors Journal **20**(12), 6762–6775 (2020)
17. Xu, X., Flynn, R., Russell, M.: Speech intelligibility and quality: A comparative study of speech enhancement algorithms. In: 2017 28th Irish Signals and Systems Conference (ISSC). pp. 1–6. IEEE (2017)