Comparative Evaluation of Speech-to-Text Software Based on Sociodemographic and Environmental Factors

Jorge Morato[1], Alejandro Pedrero[1], and Sonia Sanchez-Cuadrado[2(✉)]

[1] Carlos III University, Avda. Universidad, 30, 28911 Leganes, Spain
[2] Complutense University, C/ Santisima Trinidad 37, 28010 Madrid, Spain
sscuadrado@ucm.es

Abstract. The proliferation of voice assistants for information retrieval is propelled by technological advancements and seamless integration across multiple devices. Nevertheless, these systems face persistent limitations in accuracy and comprehension, particularly with accents, dialects, and uncommon terminology. Additional challenges include the cost of these technologies and their reliance on internet connectivity. This study conducts a comprehensive evaluation of various low-cost speech-to-text transcription software, including Windows10, Google Docs, GBoard Android, Speech-Texter, and SpeechNotes. The analysis focuses on high-error criteria in text retrieval, such as proper names, homophones, neologisms, and multilingual usage. Key variables examined include user age, message duration, and ambient noise levels. Transcription quality is meticulously assessed to determine the efficacy of voice retrieval. Results reveal significant disparities among the software, with GBoard Android demonstrating superior accuracy and the lowest error rates.

Keywords: audio information retrieval · speech recognition · speech-to-text transcription · voice assistants · Spanish · transcription quality

1 Introduction

The use of voice assistants to obtain information is continuously evolving. There are several technological reasons for this marked growth in the use of voice assistants. Some of the most significant improvements stem from advances in technology that have led to a greater ability to understand users. In this way, voice search systems offer a more convenient way to interact with technology thanks to the advancement of natural language processing and deep learning with neural networks [1, 2]. Instead of manually typing or searching for information, users can perform a query by voice command. This approach saves time and effort, which explains the current popularity of voice search systems and has even led to a shift in consumer preferences driven by the convenience and naturalness of voice interaction.

From a technical standpoint, interoperability allows voice assistants to integrate with a wide variety of applications and services, extending their functionality. These systems

T. Guarda et al. (Eds.): ARTIIS 2024 Workshops, CCIS 2349, pp. 285–299, 2025.
https://doi.org/10.1007/978-3-031-83432-5_20

are especially useful in voice-enabled devices for everyday tasks, such as smart speakers, televisions, home appliances, driving aids, smartphone searches [3], home automation, and robotics [4]. The market for speech-to-text transcription applications is expected to reach $12.1 billion by 2031 [5]. However, there are still some limitations in their performance regarding accuracy and comprehension. In practice, voice search systems apply algorithms to perform speech-to-text transcription, and it is these texts that are modeled by retrieval algorithms. Speech-to-text transcription presents various difficulties due to the inherent characteristics of speech. People speak differently depending on their dialect, accent, speech impediments, or prosody. These variations in speech make accurate transcription challenging. Another common difficulty arises from the presence of ambient or background noise. This factor can affect and interfere with the clear capture of words, leading to erroneous alterations in the transcription. Often, in the daily use of voice information programs, we encounter poor performance of some systems. The most common reason for these errors is usually due to the quality of the software and hardware used, both in the reproduction and transcription of the message. Often, in the daily use of voice information programs, we encounter poor performance of some systems. The most common reason for these errors is usually due to the quality of the software and hardware used, both in the reproduction and transcription of the message.

In this study, the evaluation will focus on the quality of transcription under parameters that are frequent sources of error in most systems. Examples of these parameters include the length of the message and the age of the speaker. It should be noted that understanding the semantics of the message, which is closely related to linguistic analysis, is excluded from the study due to its complexity. The main objective of the study is to propose an evaluation framework that can be easily adapted to the needs of specific projects. For this reason, a limited sample of both the factors studied and the subjects of the study has been chosen. A secondary objective is to design a simple method to evaluate low-cost speech-to-text software. Understanding which free software performs better has obvious advantages for users and companies with limited financial resources. This paper evaluates speech-to-text transcription in Spanish to determine the quality of speech retrieval systems. The rest of the paper is structured as follows: first, the objective of the study is outlined, followed by a presentation of the methodology used. This methodology is then applied to compare different software applications, with the results of this comparison presented in Sect. 4. Finally, some conclusions of the study are discussed.

2 Speech-to-Text Transcription

2.1 Factors Affecting Speech-to-Text Transcription

Initially, automatic speech recognition systems required high-quality audio and articulated speech. The results were limited, but accuracy rates eventually improved to surpass human performance in terms of speed and cost. Automatic speech recognition software has undergone significant evolution [6, 7], due to advancements in deep learning-based techniques [8]. Nonetheless, automatic speech recognition still presents certain imperfections [9].

For the evaluation of speech-to-text transcription, the criteria may vary. Common aspects are those related to the fidelity with the original content, whether it is the same content and meaning or the literal transcription in terms of the number of correctly transcribed terms. Grammatical consistency is also measured in terms of proper sentence structure, verb tenses, gender and number agreement and punctuation. A relevant aspect that is analyzed is the use of technical terms or specialized jargon. This terminology usually influences the quality of the transcription. Likewise, proper names must be controlled as they tend to generate errors due to their variability. Technical terminology and proper names have historically posed a challenge for natural language processing systems [10].

Apart from the characteristics of the message itself and the linguistic aspects there are other factors related to the communication channel and the characteristics of the users [11–13]. The characteristics that affect the message take into consideration the duration or clarity and quality of the audio. There are some factors that reduce the clarity of the message, such as when several speakers participate simultaneously. As for the linguistic characteristics, we observe how they affect aspects such as the simplicity and coherence of the message. The use of words from multiple languages or dialectal varieties compromises the clarity of the message [14]. Another feature that hinders the interpretation of the message is the use of homophonic words, which can lead to misinterpretation of meanings, even though the sound for transcription is the same.

On the other hand, it has been shown that ambient noise or echo during message delivery negatively affects transcription quality [15]. From the perspective of speech delivery, transcription quality can be influenced by the inherent characteristics of the user, such as age, educational level, cognitive or physical disabilities, prior knowledge of the subject, or language proficiency. Several factors contribute to the overall clarity and accuracy of a transcription:

- Message characteristics: Factors such as message length, clarity, and audio quality are crucial. For instance, the clarity of the message may diminish when multiple speakers participate simultaneously.
- Linguistic characteristics: The clarity, simplicity, and coherence of the message are essential. Challenges may arise from the use of multiple languages or regional dialects within the same message. Homophones, words with the same sound but different meanings, can also cause confusion.
- User characteristics: These include the listener's age, educational level, cognitive or physical disabilities, previous knowledge of the subject, and language proficiency.
- Communication channel: External factors such as ambient noise or echo during message reproduction can significantly affect transcription quality.

2.2 Methods for Evaluation of Speech-to-Text Transcription

To assess the quality of speech-to-text transcription, several approaches are used to measure accuracy. A common method is manual comparison, which provides precise measurements. In this approach, experts review the manual transcript, comparing the system-generated transcript with the original audio. Accuracy is assessed by identifying and correcting errors, whether they are grammatical mistakes, word omissions, or

misspellings. The main disadvantage of manual testing is that it is costly and time-consuming, which is why more and more studies are applying automatic evaluation, although not without criticism [16, 17]. A reference transcript is used, and various metrics, such as the Word Error Rate (WER), are applied to evaluate the match between the transcripts [18]. This metric measures the proportion of omitted or incorrect words in the transcript compared to the reference transcript. WER is a widely accepted metric for measuring error in word identification.

$\text{WER}\ (\%) = (S + D + I)/N.$

It uses four variables for this purpose:

(S) Number of substitutions: the word is present in the transcription but with errors.

(D) Number of deletions: words deleted in the transcription.

(I) Number of insertions, words that are in the transcript but not in the voice message.

(N) Number of words in the original text, also called reference.

Other metrics have been proposed. The Relative Information Lost (RIL) metric is based on the concept of Mutual Information (MI) and serves as a substitute for Word Error Rate (WER). It evaluates the statistical relationship between the words in the input (X) and output (Y) by utilizing Shannon Entropy H in its calculation [9]. Another approach is based on perceptual evaluation, involving a listener's judgment of the quality of the transcription. It is therefore a subjective evaluation in which evaluators rate the accuracy, comprehension, and fluency of the transcription according to a scale. It is an optimal method for capturing the quality of perception in real situations.

Corpora are often used as resources for evaluating speech-to-text transcription systems. These corpus transcriptions are typically performed by experts [19, 20]. Although there is still room for improvement in some areas [20], progress in the automatic transcription of spoken language is evident [21]. However, many cases still require manual review, as high accuracy rates do not always equate to high quality [22]. The corpora used for system evaluation should be distinct from the learning corpora to avoid influencing the results.

WER values and population samples vary across individual research studies. Pentland and colleges [16] conducted an analysis of error-prone texts derived from job interviews and manually generated transcriptions, utilizing a sample consisting of 35 male participants and 54 female participants. The researchers observed a range of WER values between 7.3 and 54.2, which were influenced by factors such as corpus characteristics, background noise, accents, and transcription quality. Iancu (2019) [23] investigated the performance of Google Cloud STT, involving 10 male participants and 10 female participants, and reported WER values ranging from 9.93 to 57.19. In a study carried out in Spanish by [24] with a sample comprising 7 male participants and 3 female participants, the author highlighted the benefits of respeaking texts as a strategy to enhance transcription accuracy. Pfeifer et al. (2024) [17] discovered that AI-generated transcripts exhibit high accuracy levels, with WER values falling within the range of 2.5 to 3.36. These researchers also observed that older adults tend to articulate more clearly and at a slower pace. The variability in WER values across studies underscores the impact of

various factors, such as speaker characteristics and linguistic diversity, on transcription accuracy.

3 Methodology

In accordance with the objectives of this study and the available resources, data were gathered from various speech-to-text transcription software applications. To compare the applications, an ad-hoc corpus containing error-prone aspects was used, and a varied population sample was selected. This work involves the following phases: 1) select message corpus; 2) select analysis parameters; 3) classify users and messages by their characteristics; 4) subjects reproduce test sentences under certain conditions; 5) define the tests and 6) the text generated by the application is subjected to evaluation. The accuracy of the result is evaluated using the WER function.

In phase 5), tests are developed at several levels for:

- Simple element or units
- Combination testing of the parameters chosen in the previous stages. In the case of the combined tests, it has been limited to the aspects of age, text size and ambient noise.
- Linguistic tests are conducted to assess aspects such as homophones, neologisms, combined languages, and proper names. The text reproduced by the application is evaluated to determine its accuracy.

We have applied the above steps to a set of five free speech-to-text software applications.

3.1 Evaluated Software

This study conducts a comprehensive evaluation of various free speech-to-text transcription software, including Windows10, Google Docs, GBoard Android, Speech-Texter, and SpeechNotes. The set of software applications studied is shown in Table 1, where the main advantages and disadvantages are listed.

- Windows 10 integrates its technology into the operating system and utilizes machine learning methods and deep neural networks. It is designed for voice recognition for dictation and system control through commands.
- Google Docs uses the Google Voice Typing system and the Google Cloud Speech-to-Text (STT) API technology. This system relies on deep neural networks (DNN) and machine learning (ML) and has been implemented for voice-to-text transcription in documents.
- GBoard (Android) employs Google's voice recognition algorithms, including deep learning models and natural language processing (NLP). It is designed to convert voice input to text on Android devices.
- Speech-Texter is based on Google's voice recognition technology, utilizing machine learning and neural networks. It has been designed for voice-to-text transcription. SpeechNotes uses Google's Speech Recognition API and employs ML and NLP models to achieve accurate voice-to-text transcription.

All these services are based on advanced voice recognition technologies, primarily supported by neural networks and machine learning. The main difference lies in how and for what purpose these technologies are used on each platform. Windows 10 offers integrated features within the operating system, while Google-based applications like Google Docs, GBoard, Speech-Texter, and SpeechNotes leverage Google Cloud infrastructure for transcription and text input. Of all the applications analyzed, the option of free use has been analyzed. The price of these applications depends on their level of use. Previously, some applications such as *Cloud Speech to Text* and *Assembly AI* were evaluated. These are highly prestigious proposals whose advanced APIs provide a high configuration capacity. However, they have been discarded for this study due to their high cost.

Table 1. Pros and cons of an initial comparison of the analyzed software.

Application	Pros	Cons
Windows10	Popularity of the operating system. Accessible design. 30 commands Languages: 8	High loading time
GoogleDocs	Endorsed by a reputable company and a popular office application. Languages: 60	Not identified
GBoard Android	Installed on mobile devices and with fast loading time Languages: 47	Not identified
Speech texter	Possibility to create your own commands Languages: 63	Poorly publicized and with little support from the community
Speech notes	Simple and usable editor	Poorly publicized and with little support from the community

3.2 Corpus for Evaluation

The message sample utilized in this study was sourced from an initiative focused on information-seeking methods targeted at Spanish-speaking undergraduate students. Participants were instructed to suggest phrases and terms they found challenging to comprehend. This method yielded a diverse, ad-hoc collection of messages, tailored to accommodate varying individual needs. Table 2 details the characteristics of the 25 participants, categorized by age and gender.

A total of 600 voice messages were recorded from the sample population, playing back the sentences under different conditions. To avoid biases related to clarity *and* comprehension, all messages were confirmed to have high readability according to the Fernandez-Huerta index and more comprehensive metrics [25]. *Sentence* complexity

Table 2. Characteristics of the sampled population in the study.

Factor	Categories	Number of Individuals
Age	Young (<10 years)	4
	Media	14
	Advanced (>60)	7
Sex	Male	10
	Female	15

levels are between very easy and normal. This prevents a higher complexity that could distort the study.

3.3 Variables

With respect to ambient noise, tests have been performed with three levels: (1) Minimum (<10 db); (2) Low (between 11–25 db); (3) Medium (26–40 db). The highest value recorded in ambient noise is 40 db, not exceeding the maximum legally allowed.

Regarding the selected criteria, different considerations on message length and linguistic aspects should be mentioned. Table 3 shows the selection of criteria and selected categories. The messages were of different lengths, classified into single word, short (less than 12 words), medium (between 13 and 30 words) or long (31–10 words) messages. Messages ranged from single words, e.g., "okay" or "hello", to long messages from literary works. Messages labeled as medium contain subordinate sentences long length corresponds to a short paragraph, which may contain several subordinate sentences.

Table 3. Examples of sentences in Spanish.

Duration	*Text*	Translation
Short	*Estos robots van a acabar dominando el mundo*	These robots will end up dominating the world
Medium	*El sol es débil cuando se empieza a elevar, y cobra fuerza y coraje a medida que avanza el día*	The sun is weak when it begins to rise, and gains strength and courage as the day progresses
Long	*Cada libro, cada volumen que ves aquí, tiene un alma. El alma de la persona que lo escribió y de aquellos que lo leyeron vivieron y soñaron con él. [….]*	Every book, every volume you see here, has a soul. The soul of the person who wrote it and of those who read it lived and dreamed with it. [….]

The sentences are labeled by collecting linguistic aspects that, according to the scientific literature, interfere with the accuracy of the transcription, such as proper names, homophonic words, terms in several languages or neologisms. Table 4 shows the criteria marked in the sentences of the message collection.

Table 4. Factors and labels used in evaluation of audio recovery systems.

Factor	Categories	Messages
Message length	Single word	2
	Short (<12 words)	2
	Medium (between 13 and 30 words)	2
	Long (between 31 and 100 words)	2
Linguistic aspects	Proper nouns	4
	Homophones	4
	Different languages	4
	Neologisms	4

For the linguistic aspects, different phrases have been defined. Examples of sentences are given below:

- Personal names example: "Dr. Fuentes was my professor." *("El doctor **Fuentes** fue mi profesor"). Rosa* is my friend ("***Rosa*** es mi amiga")
- Homophones example: If you come to live with me, you will inherit my goods ("*Si vienes a vivir conmigo, heredarás mis **bienes***"). Would you like to come for a cup of tea? ("¿***Te** vienes a tomar un **té***?). Phonetic transcription for **Te** and **té** Spanish words is /te/.
- Language mixing example: You're **cool**. ("Eres cool"), What a show! *("¡**Menudo** show!*")
- Neologisms example: If the number of cases rises, we will be confined *("Si el número de casos aumenta, estaremos **confinados**.").*

3.4 Evaluation Test

The five applications are used to evaluate each type of test designed, including unit tests, combined tests, and linguistic tests. Each software application processes eight single test sentences, considering the individualized factors outlined below. 1) Gender; 2) Age; 3) Message Length; 4) Environmental Noise; 5) Simultaneous Conversations. 6) Realtime evaluation; 7) Evaluation of orthographic signs. In the combined tests, three variables are selected for evaluation: ambient noise, message length, and the age range of the message sender. A total of 36 combinations are applied, integrating the different categories. For comparison purposes, the accuracy of the results is calculated using the WER function. Finally, linguistic tests are used to evaluate the performance of the computer tools, particularly to assess accuracy in sentences containing proper nouns, homophones, mixed languages, and neologisms.

4 Results

For the evaluation of speech-to-text transcription software, a total of 2700 analysis iterations were run (Table 5) distributed among the different types of tests. In order to avoid a large number of combinations, different test groups have been defined for each of the selected applications.

Table 5. Iterations by type of evaluation test. Each experiment is run on the 5 applications tested.

Test type	# individuals	# experiments	Total
Unitary tests	25	56 (7 tests x8 sentences)	1400
Combined test		Combines 3 different ambient noise settings with 4 dialog durations	600
Children	4	12	48
Middle age	14	12	168
Seniors	7	12	84
Linguistic tests	25	16	400
Total		108	2700

- Unit tests: each software is evaluated, for the 8 sentences, by a single factor: 1) gender; 2) age; 3) message duration; 4) ambient noise; 5) simultaneous conversations; 6) real-time evaluation; 7) evaluation with different orthographic signs. All the analyzed software has given correct results in the unit tests. That is, there were no errors in any execution.
- Combined tests: three aspects have been selected because of their weight in the tests. These criteria are described in 3.2: age, ambient noise, and message duration. Each of these criteria has different categories: for age there are 3 categories: young, middle-age, senior; for noise there are 3 categories (minimum, low and medium); and there are 4 categories according to message length, as detailed in point 3.2. All possible combinations between these categories have been performed ($3 \times 3 \times 4 = 36$ tests). The result was evaluated with the WER function.
- Linguistic tests: The results of the 16 linguistic tests are shown in the table below. These tests look at performance with homophones, proper nouns, mixed languages, and neologisms. Each of the criteria contains four test sentences, i.e. in total there are 16 sentences (4 criteria by 4 sentences). The evaluation is conducted through the accurate recognition of the problematic terms. Consequently, the findings are presented in Table 6.

The following graphs (Figs. 1 and 2) summarize the results of the combined test. The histograms show the mean results for different message lengths, for different ages and noise levels. Of note is the low error rate in middle age, as opposed to high error rates in the elderly and children.

Table 6. The linguistic performance of the software is evaluated by correctly identifying the key term of each test sentence.

Tools	Homophones	Mixed Languages	Neologisms	Proper nouns	SD
Gboard Android	3.16	3.64	3	2.88	0.29
Google Docs	2.12	3.16	2.72	3.72	0.59
SpeechNotes	3.08	3.04	3.08	3.76	0.3
SpeechTexter	2.88	2.64	3.04	3.76	0.42
W10	2.2	1.84	3.84	3.8	0.91
Average	2.688	2.864	3.136	3.584	

The analysis reveals that there is a moderate positive correlation between age and WER (0.377), indicating that the age of users has some influence on the Word Error Rate. This finding coincides with Pfeifer et al. (2024), that observes that middle aged people speak clearer than youngsters. The other factors (software, ambient noise, and duration) have very weak correlations with WER, suggesting that they have minimal impact on the error rate. The variability in WER, as indicated by the standard deviation, is moderate (14.41), and the presence of samples with no errors (minimum WER of 0) highlights the potential for achieving high accuracy under certain conditions.

The findings regarding age are more pronounced in the isolated test, where the standard deviation (SD) for the overall WER is 14.6. This SD increases to 19.85 for the elderly and 17.37 for younger individuals, while it remains at only 7.75 for the middle-aged group.

If we analyze the overall results without breaking them down, two applications stand out: Google Docs and GBoard Android. Both have a low SD (22.42 and 17.55 respectively, and an average of 30.03 and 28.08). Both applications are fast and functional, and present optimal accuracy rates and good loading times. While far from these software applications, the SpeechNotes application presents very acceptable performance. At the opposite extreme is Windows 10. W10 has the highest average WER, indicating that it tends to have more errors on average compared to the others. It also has a wide range with a high maximum WER.

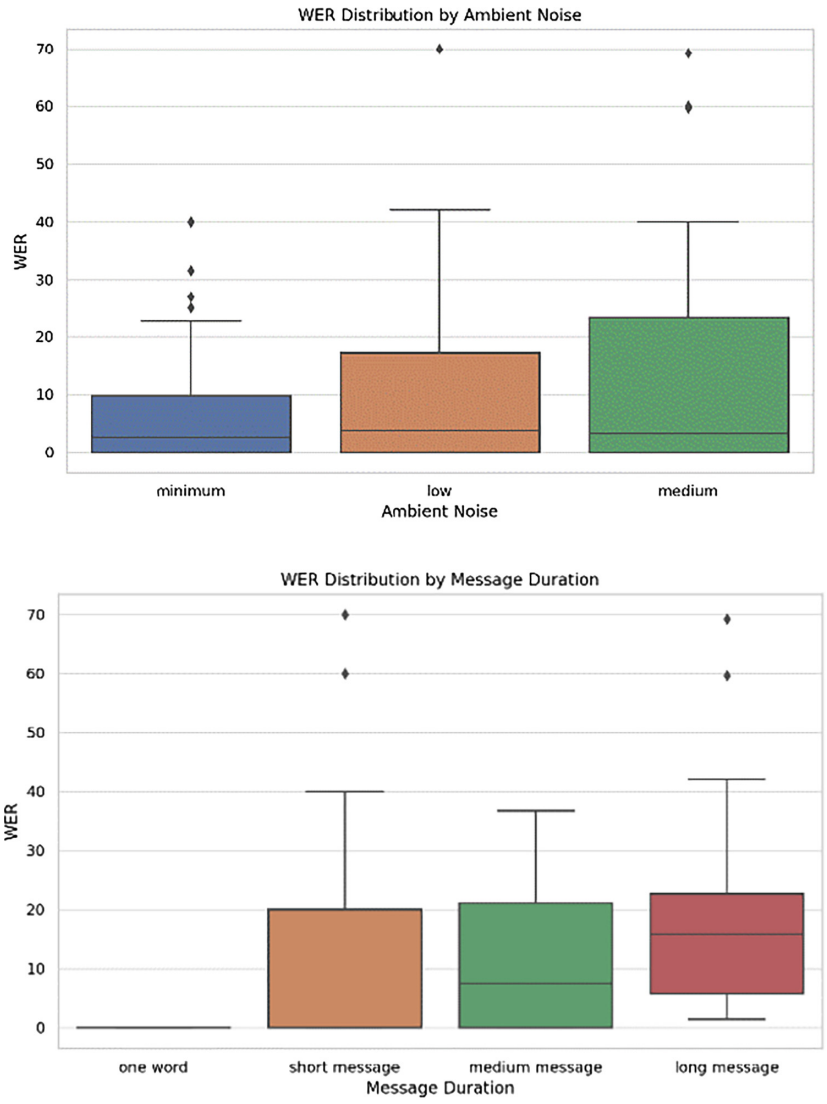

Fig. 1. Boxplot diagrams for ambient noise and message duration for the combined tests.

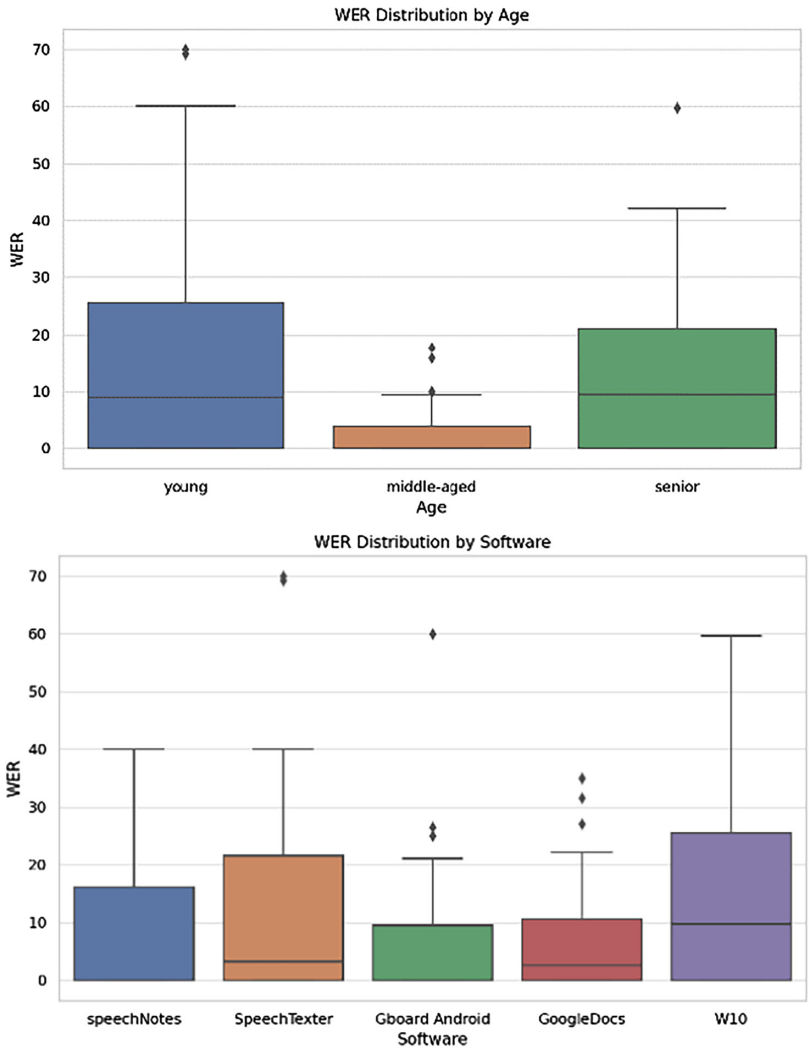

Fig. 2. Boxplot diagrams for age and Software for the combined tests.

5 Conclusions

The main findings of the study can be summarized in three aspects: 1) a straightforward methodology for the analysis of speech-to-text systems is proposed. 2) A comparison of low-cost software and freeware has been conducted under the proposed methodology; 3) Several aspects reported in the literature as a source of error have been tested with a varied sample of users. Studies on application performance are usually limited to a single aspect, either to user group characteristics (e.g. elderly or hearing impaired) or to technical aspects, such as ambient noise. Since these different aspects form a whole,

both groups have been included in the combined test. The criteria chosen were those aspects that, according to the literature, have the greatest impact.

The best performing software was *GBoard for Android*. This software is competitive with commercial alternatives. Gboard has a moderate average WER and the smallest standard deviation, suggesting more consistent performance. Although its accuracy is not 100% and it has difficulties in identifying proper names, it shows good performance and adaptability to the user.

Unit tests have shown satisfactory results. In linguistics, neologisms and language mixtures are problematic, since the language models are poorly adapted to these infrequent terms. This observation has been recently reported by authors regarding the use of ASR applied to language teaching. Perhaps the most noteworthy aspect is the relationship with age, where it seems clear that poor diction in children or physical problems in the elderly impairs the effectiveness of ASR.

Although the sample could be increased to show a more complete analysis, it should be taken into consideration that the objective of these computer applications is to give an adequate result for each individual user, something that is evidently not achieved. It should be noted that the objective in this study was not to evaluate people but applications, an objective that has been reasonably satisfied.

As for future work, we intend to increase the number of users and characteristics studied, in real situations. Many corpora, such as those from radio broadcasters, are significantly refined and do not show day-to-day situations of the general population. The analysis presented in this study was performed manually, which limits the number of aspects to be analyzed. A future development could be the automation of some aspects in order to have a more scalable methodology. The methodology depends on interviews with users, so the 2700 tests performed could be optimized with an online testing platform to reach a larger population.

Funding. Research partially funded by the R&D grant from the Autonomus Community of Madrid (PHS-2024/PH-HUM-313).

References

1. Deng, L., Liu, Y. eds: Deep Learning in Natural Language Processing. Springer Singapore, Singapore (2018). https://doi.org/10.1007/978-981-10-5209-5
2. Roger, V., Farinas, J., Pinquier, J.: Deep neural networks for automatic speech processing: a survey from large corpora to limited data. J. Audio Speech Music Proc. **2022**, 19 (2022). https://doi.org/10.1186/s13636-022-00251-w
3. Richter, F.: Infographic: Smart Speaker Adoption Continues to Rise [Infographic]. Statista Daily Data (2020). https://www.statista.com/chart/16597/smart-speaker-ownership-in-the-united-states
4. Morato, J., Sanchez-Cuadrado, S., Iglesias, A., Campillo, A., Fernández-Panadero, C.: Sustainable technologies for older adults. Sustainability **13**, 8465 (2021). https://doi.org/10.3390/su13158465
5. Kashinath, G., Kanhaiya, K., Vineet, K.: Speech-to-Text API Market. Allied Market Research, report code A09527 (2023)
6. Yu, D., Deng, L.: Automatic Speech Recognition: A Deep Learning Approach. Springer London, London (2015). https://doi.org/10.1007/978-1-4471-5779-3

7. Watanabe, S., Delcroix, M., Metze, F., Hershey, J.R. eds: New Era for Robust Speech Recognition: Exploiting Deep Learning. Springer International Publishing: Imprint: Springer, Cham (2017)
8. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**, 436–444 (2015). https://doi.org/10.1038/nature14539
9. Errattahi, R., El Hannani, A., Ouahmane, H.: Automatic speech recognition errors detection and correction: a review. Procedia Comput. Sci. **128**, 32–37 (2018). https://doi.org/10.1016/j.procs.2018.03.005
10. Marrero, M., Urbano, J., Sánchez-Cuadrado, S., Morato, J., Gómez-Berbís, J.M.: Named entity recognition: fallacies, challenges and opportunities. Comput. Stand. Interfaces **35**, 482–489 (2013). https://doi.org/10.1016/j.csi.2012.09.004
11. Humes, L.E.: Factors underlying individual differences in speech-recognition threshold (SRT) in noise among older adults. Front. Aging Neurosci. **13**, 702739 (2021). https://doi.org/10.3389/fnagi.2021.702739
12. Ogun, S.: How to create a speech dataset for ASR, TTS, and other speech tasks [Blog] (2021). https://ogunlao.github.io/blog/2021/01/26/how-to-create-speech-dataset.html
13. Tatman, R., Kasten, C.: Effects of talker dialect, gender & race on accuracy of Bing speech and Youtube automatic captions. In: Interspeech 2017, pp. 934–938. ISCA (2017). https://doi.org/10.21437/Interspeech.2017-1746
14. Winata, G.I., et al.: Learning fast adaptation on cross-accented speech recognition. In: Interspeech 2020, pp. 1276–1280. ISCA (2020). https://doi.org/10.21437/Interspeech.2020-45
15. Lu, X., Li, S., Fujimoto, M.: Automatic Speech Recognition. In: Kidawara, Y., Sumita, E., Kawai, H. (eds.) Speech-to-Speech Translation, pp. 21–38. Springer Singapore, Singapore (2020). https://doi.org/10.1007/978-981-15-0595-9_2
16. Pentland, S.J., Fuller, C.M., Spitzley, L.A., Twitchell, D.P.: Does accuracy matter? Methodological considerations when using automated speech-to-text for social science research. Int. J. Soc. Res. Methodol. **26**, 661–677 (2023). https://doi.org/10.1080/13645579.2022.2087849
17. Pfeifer, V.A., Chilton, T.D., Grilli, M.D., Mehl, M.R.: How ready is speech-to-text for psychological language research? Evaluating the validity of AI-generated English transcripts for analyzing free-spoken responses in younger and older adults. Behav. Res. **56**, 7621–7631 (2024). https://doi.org/10.3758/s13428-024-02440-1
18. Klakow, D., Peters, J.: Testing the correlation of word error rate and perplexity. Speech Commun. **38**, 19–28 (2002). https://doi.org/10.1016/S0167-6393(01)00041-3
19. Durand, J.: Corpus Phonology. In: Oxford Research Encyclopedia of Linguistics. Oxford University Press (2017). https://doi.org/10.1093/acrefore/9780199384655.013.145
20. Niemants, N.: Des enregistrements aux corpus: transcription et extraction de données d'interprétation en milieu médical. meta. **63**, 665–694 (2019). https://doi.org/10.7202/1060168ar
21. Ravanelli, M., Brakel, P., Omologo, M., Bengio, Y.: Light gated recurrent units for speech recognition. IEEE Trans. Emerg. Top. Comput. Intell. **2**, 92–102 (2018). https://doi.org/10.1109/TETCI.2017.2762739
22. Dias, G.: Dossier: IA & technologies du langage humain. Bulletin de l'AFIA **107**, 6–9 (2020)
23. Blackley, S.V., Huynh, J., Wang, L., Korach, Z., Zhou, L.: Speech recognition for clinical documentation from 1990 to 2018: a systematic review. J. Am. Med. Inform. Assoc. **26**, 324–338 (2019). https://doi.org/10.1093/jamia/ocy179
24. Iancu, B.: Evaluating Google Speech-to-Text API's performance for Romanian e-learning resources. Informatica Economica **23**(1), 17–25 (2019). https://ideas.repec.org/a/aes/infoec/v23y2019i1p17-25.html

25. Rufino Morales, M.: Estudio comparativo de métodos de transcripción para corpus orales: el caso del español. Revista Nebrija de Lingüística Aplicada a la Enseñanza de Lenguas **14**, 126–146 (2020). https://doi.org/10.26378/rnlael1429406
26. Serna, Y., Morato, J, Sanchez-Cuadrado, S.: Evaluación de la comprensión de los paneles interpretativos en parajes naturales. Scire **24**, 53–62 (2018). https://doi.org/10.54886/scire.v24i2.4568

Online Resources

27. AssemblyAI: The #1 Speech-to-Text API for Developers. https://www.assemblyai.com/
28. Cloud Speech-to-Text API - Marketplace - Google Cloud Platform. https://console.cloud.google.com/marketplace/product/google/speech.googleapis.com
29. Google Documents: create and edit documents online for free. https://www.google.es/intl/es/docs/about/
30. Gboard: Google's keyboard - Apps on Google Play. https://play.google.com/store/apps/details?id=com.google.android.inputmethod.latin&hl=es&gl=US
31. Speech to Text Online Notepad. Free. Speechnotes. https://speechnotes.co/
32. SpeechTexter. Type with your voice online. Speech Texter. https://www.speechtexter.com