




Detecting Speech Disfluencies Using Open-Source Tools in Automatic Feedback Systems for Oral Presentation Training

Willy Mateo²^a, Leonardo Eras¹^b, Giancarlo Carvajal²^c and Federico Domínguez^{1,2}^d

¹Information Technology Center, Escuela Superior Politecnica Del Litoral, ESPOL, Guayaquil, Ecuador

²Faculty of Electrical and Computer Engineering, Escuela Superior Politecnica Del Litoral, ESPOL, Guayaquil, Ecuador
{wjmateo, leras, gianflor, fexadomi}@espol.edu.ec

Keywords: Oral Presentation Skills, Filler Words Detection, Filled Pauses Detection, Automatic Presentation Feedback.

Abstract: In the realm of verbal communication, most common non-clinical speech disfluencies are filler words and filled pauses, which pose challenges for effective oral presentations. Yet their detection is no easy task. This article presents the usage of OpenAI's Whisper for filled pauses and filler words detection in Spanish oral presentations, including on-the-wild usage with undergraduate students. Preliminary results indicate that Whisper demonstrates promise as a valuable tool to identify a substantial amount of filler words and filled pauses. Despite areas of improvement, Whisper serves as a diagnostic tool for assessing disfluencies in oral communication.

1 INTRODUCTION


Speech disfluencies are interruptions in the regular flow of speech (Scott Fraundorf, 2014). These interruptions may manifest in several forms during speech ranging from mild interruptions such as filler words, where the disfluent sound is a complete word (e.g. *ok, like*), and filled pauses, where a natural pause during speech is vocalized with prolonged vowels (e.g. *uhmm, ehmmm*), to speech disorders such as stuttering (Gósy, 2023; Das et al., 2019). While some moderate use of filled pauses and filler words may in fact increase listener comprehension and add naturalness to speech (Scott Fraundorf, 2014), it is commonly understood that increased usage of these disfluencies denote hesitation in the speaker and delay comprehension in the listener (Lo, 2020).


Practice, self-awareness, and feedback are frequently cited as strategies to reduce the occurrence of speech disfluencies in oral presentations (De Grez et al., 2009; Alwi and Sidhu, 2013; Das et al., 2019). Almost all coaching systems for oral presentations attempt to detect speech disfluencies either in real time or offline from recordings to provide automatic feedback to learners. However, obtaining reasonable


levels of detection accuracy is no easy task. Some systems use phonetic analysis with software such as Praat (Boersma and Weenink, 2023), a well-tested, production-ready open-licensed software for speech analysis and phonetics, to extract filled pauses from speech, while others use Automatic Speech Recognition (ASR) to identify filler words or repetitions from generated transcripts (Zhu et al., 2022). Phonetic analysis excels at detecting filled pauses but struggles with repetitions while ASR technologies, normally trained to ignore disfluencies, typically require costly annotated data for retraining (Zhu et al., 2022).


In this work, we focus on the detection of the most common non-clinical speech disfluencies: filler words and filled pauses (Lo, 2020), using ASR technologies with open source tools. Specifically, we tackle on the problem on providing accurate detection of speech disfluencies in Spanish speaking students in the context of an oral presentation.

Since 2018 we have been using the Automatic Feedback Presentation system (RAP for its Spanish acronym) as an experimental tool that facilitates learning of oral presentation skills and since 2019 as a learning tool embedded in the academic activities of communication courses and selected engineering courses (Domínguez et al., 2021). The RAP system records a student's oral presentation in a specialized room and extracts five presentation features: posture, gaze to the audience, use of filled pauses, voice volume, and slides legibility (Ochoa et al., 2018). The

^a <https://orcid.org/0009-0003-0616-7717>

^b <https://orcid.org/0000-0002-3594-9289>

^c <https://orcid.org/0009-0003-0664-7622>

^d <https://orcid.org/0000-0002-3655-2179>

system uses Praat to measure voice volume and detect filled pauses from the audio recording. It uses this extracted features to generate an interactive and customized feedback report at the end of the presentation. In 2020, we demonstrated that this facilitates modest, but statistically significant, learning gains to their users (Ochoa and Dominguez, 2020).

Aiming to improve the capacity of the RAP system to detect speech disfluencies, we implemented two detection algorithms, based on OpenAI’s Whisper (Radford et al., 2022), for filler words and filled pauses. Section 2 summarizes the state of the art on the detection of non-clinical speech disfluencies, section 3 describes the experimental methodology for evaluation of both algorithms with annotated data and in a small on-the-wild setup within the RAP system, section 4 describes the obtained results, and sections 5 and 6 summarize our findings.

2 RELATED WORK

The domain of identifying and classifying filler words in audio recordings is currently evolving and ongoing. A relevant study, conducted in 2022 (Zhu et al., 2022), focused on the detection and classification of filler words in English audio recordings. To do this, audio podcasts collected from the online music platform SoundCloud were used, with a total of 145 hours of recordings. The focus of this research was based on the creation of a pipeline using Voice Activity Detection (VAD) and ASR techniques, which allowed a more accurate and efficient identification of filler words present in recordings compared to the keyword-based approach.

More recent work, conducted in June 2023 (Zhu et al., 2023), addresses the challenge of detecting non-linguistic filler words. It is recognized that universal accessibility to ASR systems may be limited by factors such as budget constraints, the diversity of target languages, and the computational resources required. Through the employment of Structured State Space Sequence (S4) models and semi-Markov neural conditional random fields (semi-CRF), an absolute performance improvement of 6.4% at the segment level and 3.1% at the event level is achieved in the Podcast-Fillers dataset.

As for commercial tools, Microsoft’s ability to evaluate oral presentations through the Speaker Coach (Microsoft, 2023) tool within the PowerPoint application stands out. This tool allows simulations of oral presentations, providing detailed results on the quality of the presentation. It evaluates various parameters such as speech rate, pronunciation, tone of

voice, repetitive language, inclusive language, originality, and use of filler words. However, this functionality is restricted to the Microsoft PowerPoint environment, lacking an API to integrate it into external projects and it does not offer the possibility to download a detailed report of the evaluation results.

In addition, various web-based commercial options offer similar functionalities, each with specific approaches and payment plans:

- Kapwing (Eric and Julia, 2023) introduces the Smart Crop tool, which detects filler words allowing users to remove these audio or video segments.
- Podcastle (Team, 2022) offers the Filler Word Detection tool which provides the audio transcription, detecting filler words, and allowing users to delete the specific segment.
- Cleanvoice AI (Adrian, 2023) specializes in removing unnecessary sounds, such as filler sounds, stuttering and mouth noises.

3 METHODOLOGY

The tools described in the previous section are either experimental or commercial. Aiming to integrate our findings in the RAP system, we focused on open-source solutions.

3.1 Filler Words

For detecting filler words out of speech transcriptions using Speech to Text (STT) technology, two open-source solutions stand out: Whisper and Coqui STT.

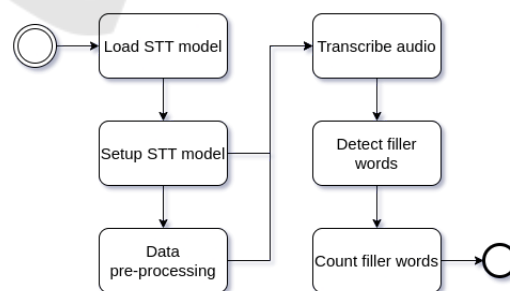


Figure 1: Process of filler word detection.

Whisper (Radford et al., 2022) is a model specialized in speech processing including multilingual speech recognition, speech translation, spoken language identification and voice activity detection. It has an architecture based on Transformer sequence-to-sequence and was trained on an extensive dataset comprising 680 000 hours of audio in 97 languages,

including Spanish. This training resulted in the creation of 5 models variants, presented in ascending order from lowest to highest capacity: Tiny, Base, Small, Medium and Large.

The Medium model was used to identify filler words. There was no need to apply conversion or cleanup techniques to the audio as Whisper internally splits the audio conveniently to deliver optimal results.

3.1.1 Detection and Counting of Filler Words

Previously, communication courses teachers from Escuela Superior Politécnica del Litoral (ESPOL®) shared with us a list of filler words most frequently used by students in an oral presentation. With this information as a foundation, a dictionary was built with common fillers words as keys and a list with the variations of said filler word as values.

We noticed that when using Coqui STT it was difficult to detect certain filler words, so, a boost was added to those words. In addition, it was unable to detect composed filler words, for example: "cómo es que es", limiting its applicability.

At this point, we proceeded to either generate the transcription of all the audio chunks for Coqui STT, or to process the entire audio and obtain the transcription of each segment if Whisper was used.

To save the the number of filler words detected in an audio file it uses JSON files, this is an open standard file format used to store and transmit data objects consisting of attribute–value pairs and arrays. A JSON file is created with filler words as keys, with all values set to 0, representing the number of times each filler word was detected. Additionally, another JSON is initialized with the same keywords, assigning empty arrays as values that will store the specific moments in which each filler word was uttered.

Next, the dictionary of common filler words is iterated in descending order by the number of words in each filler word. In each iteration, by means of the `nlk`¹ python library focused on NLP (Natural Language Processing), the transcription tokens are extracted to create a list of n-grams with size equal to the number of words and obtain their frequency distribution, this is useful to count the number of occurrences of the filler word and its variations in the transcription (Bird et al., 2009).

The JSON is then updated with the number of filler words detected. In the case of Whisper, the start and end seconds when each filler word was detected, obtained from the corresponding metadata, are incorporated. Finally, the filler word and its variations

are removed from the transcription. This procedure avoids consideration of these words in future iterations, preventing them from being intertwined with terms from different filler words.

3.1.2 Tool Comparison and Performance Analysis

To calculate the performance of each STT technology to detect filler words, it was necessary to perform the manual transcription of each test audio.

We tailored a custom metric called Filler Word Error Rate (FWER), based on the Word Error Rate (WER) formula which is a commonly used measure in the evaluation of speech recognition or machine translation systems. FWER takes into account both filler word insertions and deletions in the output of STT technology compared to the total number of filler words in the manual transcription. So, we calculated the FWER for each of the test audios using both Whisper and Coqui STT.

$$FWER = \frac{\text{insertions} + \text{deletions}}{\text{filler words in manual transcript}}$$

3.2 Filled Pauses

To correctly capture filled pauses, we rely on the definition presented by (María J. Machuca, 2015), which states that the /e/ and /m/ phonemes are the only ones that can constitute true filled pauses in Spanish oral speeches. We then define a filled pause as a word that doesn't belong to the speech and fits one of the /e/, /a/ or /m/ phonemes. We complete this definition by adding prolongations, which are the prolonged sounds of a vowel (and sometimes a consonant) at the end of a word. If these prolongations are long enough, they effectively function as filled pauses. From now on, the term filled pause will refer to this specific combined definition.

3.2.1 Whisper Setup

Our filler word detection algorithm relies on the Whisper Small model, rather than the Medium one. While we initially used both models, we found that the Small model consistently produced more accurate results in several comparisons. As a result, we made the decision to exclusively use the Small model going forward. We used the `transcribe` function, with the `temperature` value set to 0 to make the transcription process deterministic, and a string in the `initial_prompt` parameter.

¹<https://github.com/nltk/nltk>

3.2.2 Detection and Counting of Filled Pauses

By default, Whisper models may leave out filled pauses, so we used **prompting**² to avoid this. The presence of ellipses (...) plays a very important role in prompting. The following string is passed in the `initial_prompt` parameter because it showed the best results:

```
Eh... Rust es... un lenguaje
eh... de programación eh... que
ah... compite eh.. directamente
con... ah... C
```

We split the original audio in 10 seconds clips. For each clip, we call the `transcribe` function and iterate through every word of the transcript, using the following regular expression to capture filled pauses:

```
.{1,2}[hm]{1,2}\.\.\.
```

It fits the first part of the provided filled pause definition and effectively matches strings as “eh...”, “ah...”, “mmm...”, which represent common Spanish filled pauses.

For the second part of the definition, we use this regular expression, as the ellipses represent prolongations themselves:

```
\.\.\.
```

Due to the fact that intonation influences whether a prolongation is considered a filled pause and that we do not analyze the audio itself, but the transcript provided by Whisper, we need to minimize the presence of false positives when capturing prolongations. We achieve this by adding a time threshold of **0.95** seconds to make sure that the prolongations are true filled pauses most of the time.

When a filled pause (fp) is detected, two discard conditions are checked, due to the possibility of the models entering a repetition loop and transcribe a word more than once. If the detected filled pause meets either of them, we ignore it:

- The detected fp has the same end timestamp than the previous one
- The detected fp is identical to the previous one

Finally, we increase the fp counter and use the word timestamps to export an audio clip containing the filled pause with `pydub`.

²<https://platform.openai.com/docs/guides/speech-to-text/prompting>

3.2.3 Performance Analysis

Since we only rely on accuracy and recall, we use the F-score and the average of false positives as the performance analysis metrics for this algorithm. We also calculate these metrics for the existing Praat-based filled pause detector to compare.

3.3 Processing Pipeline of Speech Disfluencies Detection in the RAP System

The RAP system is designed to efficiently handle media recordings and presentation slides. Once a presentation is completed, the system bundles the media recordings, presentation slides, and metadata into a compressed zip file. This file is then transmitted over the network to a backend server repository using a reliable file transfer protocol. The backend server uses a broker that implements a publish-subscribe message protocol to facilitate communication and orchestration between all the components of the RAP system. Upon arrival at the backend, the broker notifies a file-consumer service, which decompresses the zip file and notifies other media-specific services that the media files are ready for processing. This entire process is managed through the broker, ensuring seamless coordination between all the components of the RAP system.

Two services process the audio file sequentially: the first one to detect and count filler words and generate a bar chart; the second one to detect and count filled pauses.

Finally the generated bar chart, along with the audio snippets, are sent to a web service where it becomes part of a report for both the student and the professor. Figure 2 illustrates the overall process previously described. Figure 3 illustrates the audio file process previously described.

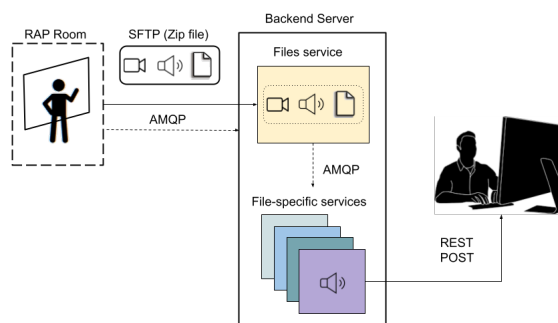


Figure 2: Complete process of the audio file from the recording room to the analyzers and finally to become part of the presentation report.

3.4 On-the-Wild Evaluation Setup

As described in the previous section, we added the detection of filler words as an experimental feature in the RAP system to evaluate this feature “on-the-wild”. This implied that after a RAP presentation, an additional bar chart, together with a small explanation about filler words, was added to the student’s RAP report. Figure 4 depicts an example of a bar chart, presented in a student’s RAP report, with the most common filler words used by the student during their presentation. The chart was only informative, no evaluation of the use of filler words was presented to the student.

All students were informed during class sessions about the usage of a new experimental feature and presented with an online informed consent form explaining the nature of the experiment and how their data will be handled. Consent was completely optional with no effect to their academic activities and could be withdrawn at any time by the students. Consent implied using the student’s data (academic metadata and RAP results) for this research in an anonymized aggregated form. We contrasted the students’ RAP results against three variables: gender, study program, and academic performance. Gender is self-reported by the student, study program and academic performance (grade average) are part of the student’s academic metadata.

In total, 290 students from three different courses gave their informed consent (90% of enrolled in these courses) to participate with their data on this research. Specifically, 211 from the course *Communication*, 43 from the course *Embedded Systems*, and 49 from the course *Computation and Society*. Communication students used the RAP system twice in the semester, the rest used the system only once. Some Communication students (13) were also enrolled in *Computation and Society* and consequently used the RAP system three times. Overall, 501 presentation recordings were obtained from participants.

Additionally, we performed a post hoc detection of filled pauses on the 501 recordings using the Whisper methodology with the objective of evaluating its effectiveness on real RAP presentations. These results were not presented in the student’s RAP report because they were obtained several weeks after their presentations.

4 RESULTS

To measure the behavior of both STT technologies (Whisper and Coqui STT), an analysis of 64 au-

dios was carried out, all recorded in the RAP classroom and each with an average duration of 5 minutes. These standardized conditions ensured consistency in data collection and provided a controlled scenario for our assessments.

Table 1: FWER percentage by STT technology.

STT technology	FWER [%]
Whisper	35.85
Coqui STT	90.80

Table 1 summarizes the error rate of filler words detection. There is a significant disparity in the accuracy of both technologies. According to Errattahi, the WER is a good metric to know which ASR technology is better than another (Errattahi et al., 2018); taking into account that the FWER metric was based on the WER and that the filler word error rates of Whisper and Coqui STT technologies are 35.85% and 90.80% respectively, we can state that Whisper is better than Coqui STT for detecting filler words in oral presentations. This may be because most of the dictionary of common filler words were added as hot words for Coqui STT.

Table 2: FWER percentage by Gender using Whisper.

Gender	Num. Audios	FWER [%]
Women	21	27.47
Men	43	26.87

Table 2 shows interesting patterns in the detection of filler words according to the gender of the speaker using Whisper. The FWER for women’s presentations is 27.47%, while for men’s presentations it is 26.87%. These results suggest a consistency in the effectiveness of Whisper in the detection of filler words regardless of the gender of the speaker and indicates that there is no gender bias in the detection of filler words. As there might be gender differences in the use of speech disfluencies, it is important to rule out gender bias in the detection technology.

Regarding filled pauses, 55 of these same 64 audios were used. There were 694 real filled pauses. We listened each detected filled pause clip, for both Whisper and Praat implementations, to count true and false positives and used manual transcripts for the false negatives.

Table 3: Performance metrics of filled pause detection.

Implementation	F-score	FPS Avg.
Whisper	0.8816	1.29
Praat	0.3541	17.09

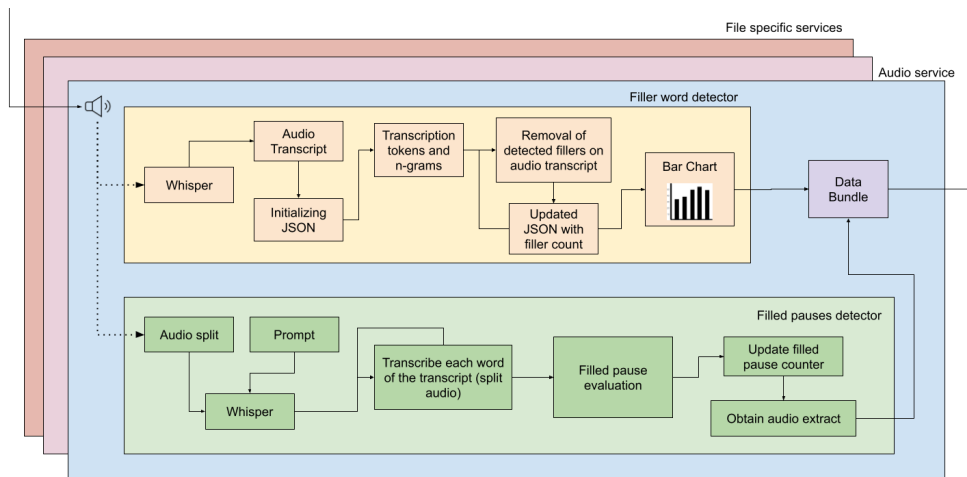


Figure 3: Phases of audio evaluation to obtain both filler words and filled pauses.

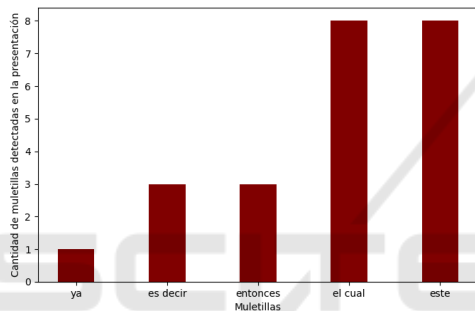


Figure 4: Example of a report of frequency of usage of filler words in a five-minute presentation.

Table 3 presents the obtained F-scores and false positives averages for the two implementations of the filled pause detection algorithm.

Table 4: Expanded performance metrics of filled pause detection.

Impl.	TP	FN	FP	Accuracy	Recall
Whisper	603	91	71	0.9120	0.8689
Praat	316	775	378	0.3247	0.4553

Table 4 collects the individual calculated metrics. The accuracy value is the weighted mean of the accuracy of each recording analysis.

4.1 On-the-Wild Evaluation

Some recordings, due to unexpected background noises and technical errors, were not processed (488 out 501 were processed for filler words, 481 out of 501 were processed for filled pauses). Use of the RAP system was divided in two sessions during the semester, Table 5 explains the use of the RAP system by subject.

Table 5: RAP sessions by subject used in the evaluation.

Subject	Session 1	Session 2
Communication	YES	YES
Embedded Systems	YES	NO
Computation and Society	NO	YES

Figure 5 shows the filler words results by gender (women = 109, men = 177) and Figure 6 by study program. It can clearly be seen that, on average as reported by the mean and the median, men (mean = 12.9) use more filler words than women (mean = 8.9) in a five-minute oral presentation. As data is non-normally distributed, a Wilcoxon non-paired two-sample t-test was performed on results on the first and second sessions. On both cases, highly statistical differences were observed. As for study programs, statistically significant differences (using the same test) were observed between engineering programs and non-engineering programs. For example, on average, Electronics and Computing engineering students reported higher (mean = 13.8) use of filler words than Design and Communication students (mean = 8.6). Statistically significant differences remained, to a lesser degree due to diminished sample size, when controlling by gender and voice volume. No statistically significant differences were observed by academic performance. No statistically significant differences were observed for Whisper filled pauses on any variable.

Figure 7 shows the Pearson correlation statistics between RAP results and Whisper filled pauses and filler words. While all correlations were either non-existent or weak, a statistically significant positive correlation ($r \neq 0$) was observed between filler words and filled pauses (as extracted by Praat) and a negative correlation between filler words and gaze. As ex-

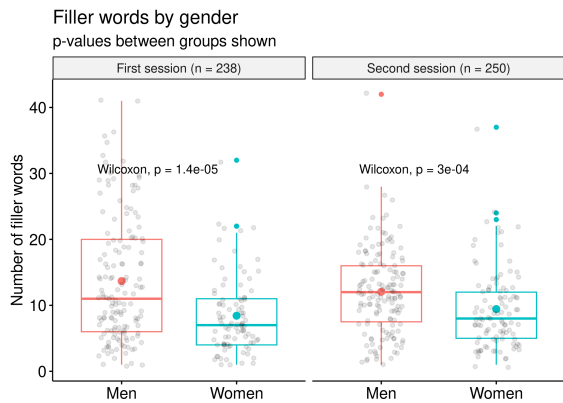


Figure 5: Use of filler words by gender in RAP presentations in the first and second sessions.

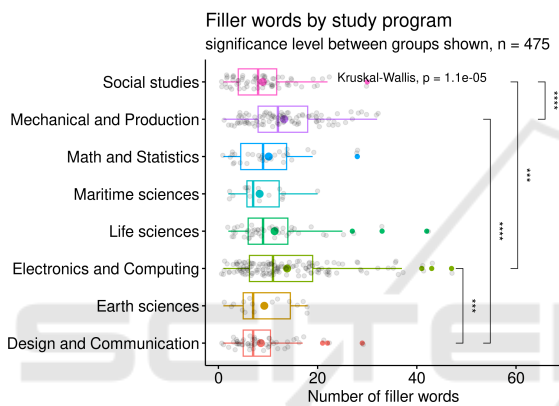


Figure 6: Use of filler words by study program in RAP presentations, both sessions.

pected, a positive, but rather weak, correlation is observed between both filled pauses (Whisper and Praat) detection algorithms.

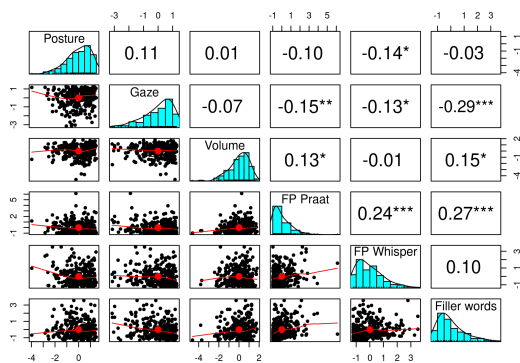


Figure 7: Correlation between RAP scores and Whisper-extracted filler words and filled pauses.

5 DISCUSSION

Whisper, with its medium model and ability to transcribe audios in Spanish with 3.6% WER, offers an opportunity to apply it to the identification of filler words in oral presentations. The evaluation of Whisper technology in the context of the RAP system shows a FRER error rate of 35.85%, indicating a moderate identification of filler words that can improve feedback to students in an academic context. This efficacy is especially significant considering the complexity of identifying filler words in oral presentations.

The results of the on-the-wild evaluation corroborate a pattern that has been observed before in language research: there is a gender disparity in the use of non-clinical speech disfluencies (Bortfeld et al., 2001; Lo, 2020). On average, men tend to use ~40% more filler words than women in a five-minute RAP presentation. It is important to point out that the recordings of male presenters tend to exhibit slightly higher voice volume levels, which may facilitate the detection of disfluencies. However, after controlling for voice volume, the gender effect remains. After controlling for study program, there are more men in engineering programs, the gender effect still remains, albeit with a weaker statistical significance due to reduced sample size.

As for the poor correlation between both the Praat and Whisper filled pauses detection algorithms, it might happen that the latter needs further improvement or that the large number of false positives in the former interferes. Nevertheless, Whisper results are quite decent, clearly outperforming the existing Praat algorithm as it is shown in the results tables. Whisper small model should certainly be the first choice for detecting filled pauses, before trying to fine-tune a model or creating a new detector from scratch.

In previous work, we found that engineering students tend to do poorly with non-verbal oral presentation skills, as measured by posture and gaze to the audience (Domínguez et al., 2023). It is interesting to note that this pattern is also observed in verbal skills as measured by filler words. Also, the inverse correlation ($r = -0.29$, highly significant), observed between gaze to the audience and filler words suggest a possible correlation between verbal and non-verbal oral presentation skills.

6 CONCLUSIONS

This study presents the evaluation of different tools to detect both filler words and filled pauses. For filler

words, Whisper outperforms Coqui STT, while providing similar performance across gender disparities. For filled pauses, the small Whisper model provides a balance between good accuracy and recall, compared to the medium model.

While the Whisper algorithms have room for improvement, the tool performs well on-the-wild and may allow the exploration of possible correlations between verbal and non-verbal oral presentation skills.

ACKNOWLEDGEMENTS

We express our sincere gratitude to Maria Gonzalez, Nicole Asqui and Hayleen Carrillo for their invaluable collaboration in the execution of this project. We also extend our appreciation to all the contributors to the open source libraries and tools used in its development.

REFERENCES

- Adrian, S. (2023). Filler words remover. Accessed: 2023-12-29.
- Alwi, N. F. B. and Sidhu, G. K. (2013). Oral Presentation: Self-perceived Competence and Actual Performance among UiTM Business Faculty Students. *Procedia - Social and Behavioral Sciences*, 90(InCULT 2012):98–106.
- Bird, S., Klein, E., and Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit*. "O'Reilly Media, Inc."
- Boersma, P. and Weenink, D. (2023). Praat: doing phonetics by computer. <http://www.praat.org/>. Accessed: 2023-12-19.
- Bortfeld, H., Leon, S., Bloom, J., Schober, M., and Brennan, S. (2001). Disfluency Rates in Conversation: Effects of Age, Relationship, Topic, Role, and Gender. *Language and Speech*, 44(2):123–147.
- Das, S., Gandhi, N., Naik, T., and Shilkrot, R. (2019). Increase apparent public speaking fluency by speech augmentation. In *2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Brighton, UK. IEEE.
- De Grez, L., Valcke, M., and Roozen, I. (2009). The impact of goal orientation, self-reflection and personal characteristics on the acquisition of oral presentation skills. *European Journal of Psychology of Education*, XXIV:293–306.
- Domínguez, F., Eras, L., Tomalá, J., and Collaguazo, A. (2023). Estimating the Distribution of Oral Presentation Skills in an Educational Institution: A Novel Methodology. In *International Conference on Computer Supported Education, CSEDU - Proceedings*, volume 2, pages 39–46. SCITEPRESS.
- Domínguez, F., Ochoa, X., Zambrano, D., Camacho, K., and Castells, J. (2021). Scaling and Adopting a Multimodal Learning Analytics Application in an Institution-Wide Setting. *IEEE Transactions on Learning Technologies*, 14(3):400–414.
- Eric, L. and Julia, E. (2023). Remove filler words. Accessed: 2023-12-29.
- Errattahi, R., El Hannani, A., and Ouahmane, H. (2018). Automatic speech recognition errors detection and correction: A review. *Procedia Computer Science*, 128:32–37. 1st International Conference on Natural Language and Speech Processing.
- Gósy, M. (2023). Occurrences and Durations of Filled Pauses in Relation to Words and Silent Pauses in Spontaneous Speech. *Languages*, 8(1).
- Lo, J. J. (2020). Between Äh(m) and Euh(m): The Distribution and Realization of Filled Pauses in the Speech of German-French Simultaneous Bilinguals. *Language and Speech*, 63(4):746–768.
- María J. Machuca, Joaquim Llisterri, A. R. (2015). Las pausas sonoras y los alargamientos en español: Un estudio preliminar. *Revista Normas*, 5:81–96.
- Microsoft, T. (2023). Rehearse your slide show with speaker coach. Accessed: 2023-12-29.
- Ochoa, X. and Dominguez, F. (2020). Controlled evaluation of a multimodal system to improve oral presentation skills in a real learning setting. *British Journal of Educational Technology*, 51(5):1615–1630.
- Ochoa, X., Domínguez, F., Guamán, B., Maya, R., Falcones, G., and Castells, J. (2018). The RAP System : Automatic Feedback of Oral Presentation Skills Using Multimodal Analysis and Low-Cost Sensors. In *LAK'18: International Conference on Learning Analytics and Knowledge*, pages 360–364, Sydney, Australia. ACM.
- Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., and Sutskever, I. (2022). Robust speech recognition via large-scale weak supervision.
- Scott Fraundorf, Jennifer Arnold, V. L. (2014). Disfluency. obo in linguistics.
- Team, P. (2022). Improve the way you sound! remove filler words from text in seconds! Accessed: 2023-12-29.
- Zhu, G., Caceres, J.-P., and Salamon, J. (2022). Filler word detection and classification: A dataset and benchmark. *arXiv preprint arXiv:2203.15135*.
- Zhu, G., Yan, Y., Caceres, J.-P., and Duan, Z. (2023). Transcription free filler word detection with neural semi-crfs. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.