# Automatic speech recognition (ASR) with Whisper: Testing Performances in Different Languages.

Terry Amorese[1,*,†], Claudia Greco[1,†], Marialucia Cuciniello[1,†], Rosa Milo[1,†], Olga Sheveleva[1,†] and Neil Glackin[2,†]

[1]*Dipartimento di Psicologia, Università degli Studi della Campania "Luigi Vanvitelli", Viale Ellittico, 31 - 81100 Caserta, Italy.*

[2]*Intelligent Voice Ltd., St Clare House, 30-33 Minories, London, United Kingdom*

**Abstract**

The present work aims at testing transcription performances of Whisper, an Automatic Speech Recognition (ASR) model produced by Open AI. The tool has been adopted in the context of a research project aimed at developing an automatic depression diagnosis support system, in order to transcribe audio data obtained from depressed and healthy subjects belonging to different countries (UK, Italy, Russia).

Results showed optimal Whisper performances for each language in terms of Number of correct words (COR) recognition, with low error rates concerning the Number of deleted words (DEL), Number of substituted words (SUB), Number of inserted words (INS) e truly low Word Error Rates (WER).

**CCS CONCEPTS**

Human-centered computing, Collaborative and social computing, Collaborative and social computing design and evaluation

**Keywords**

Automatic Speech Recognition, Depression, Whisper

## 1. Introduction

Considered the growing presence of technology in our personal and work life, it is inevitable to think about how it also affects the way we communicate and collaborate with others, as well as the positive influence that technologies as Interactive Artificial Intelligence could have in supporting mental health.

The work presented in this paper born in the context of a research project called "Androids (AutoNomous DiscoveRy of Depressive Disorder Signs)" which is aimed at developing an automatic depression diagnosis support system. The aim is to identify features of speech expressions that may signal the presence of a depressive state. More specifically, we focused on verbal behavior analysis, which allows the investigation of both the content (what is said, i.e.,

*S3C'23: Sustainable, Secure, and Smart Collaboration Workshop, Hosted by CHITALY 2023, SEPTEMBER 20–22, 2023, TURIN, ITALY*

*Corresponding author.

†These authors contributed equally.

terry.amorese@unicampania.it (T. Amorese); claudia.greco@unicampania.it (C. Greco); marialucia.cuciniello@unicampania.it (M. Cuciniello); rosa.milo@unicampania.it (R. Milo); olga.sheveleva@unicampania.it (O. Sheveleva); neil.glackin@intelligentvoice.com (N. Glackin)

the nouns) and the linguistic style (how it is said, e.g. pronouns, prepositions, articles) [11]; two aspects which could reflect cognitive patterns (e.g. self-focus, pessimism, low self-esteem [3, 12], and emotional states (e.g. anger, anxiety, sadness) that are dominant and/or maladaptive in depressive disorders.

The goal is to investigate the speech content of depressed clinical group through a computerized text analysis using the Linguistic Inquiry Word Count [LIWC] tool [11] which classifies words of a given text into several categories. However, to be able to do this work we had to first transcribe the collected audio data, using Whisper, an Automatic Speech Recognition (ASR) model produced by OpenAI [9].

Typically, ASR systems receive acoustic input from a speaker through a microphone, analyze the input through pattern, model, or algorithm, and produce an output, usually in the form of a text [6]. The accuracy of the performance of a speech recognition system is affected by many factors, as for instance the dependence or independence from the speaker, the discrete or continuous modality of word recognition, the vocabulary, and the environment [1]. Whisper is the latest in a series of Convolutional Transformer End-to-End ASR models, similar in structure to Wav2Vec2 [2] and WavLM [4].

However, what sets it apart is the scale of the training of the model. Whisper was trained on 680.000 hours of multilingual audio data with transcription collected from the web. Ground truth transcriptions for the audio training data were collected from subtitle files accompanying the audio and filtered using various heuristics in an attempt to ensure ground truth quality, while striving for dataset scale. The resulting approach is termed weakly supervised, as each ground truth transcription is not human reviewed but is instead filtered automatically. The scale of the model and the effectiveness of the weak supervision have seen Whisper exceeding previous state of the art benchmarks for multilingual ASR, notable particularly as training sets of said benchmark corpora are not included in Whisper training data.

In this work, in which we use the pretrained Whisper Medium model for all of the automated transcription, will be showed Whisper performances on audio data collected in different countries (United Kingdom, Italy and Russia) both with healthy participants and clinical groups leaving with depressions.

## 2. Methodology

Whisper's performances were tested on a sample of 226 participants, split into three groups:

Group 1: 65 English participants, 30 healthy subjects with no history or current conditions of any psychiatric disorders (16 females and 14 males, mean age = 50.9; Standard Deviation = ±11.7) and 35 patients diagnosed with Major depressive disorder (17 females and 18 males, mean age = 45.5; Standard Deviation= ±13.9).

Group 2: 93 Italian participants, 49 healthy subjects with no history or current conditions of any psychiatric disorders (40 females and 9 males, mean age = 47.5; Standard Deviation = ±11.6) and 44 patients diagnosed with Major depressive disorder (34 females and 10 males, mean age = 44.4; Standard Deviation = ±12.9).

Group 3: 68 Russian participants, 34 healthy subjects with no history or current conditions of any psychiatric disorders (18 females and 16 males, mean age = 17; Standard Deviation = ±1.7) and 34 subjects leaving with depressive disorders (20 females and 14 males, mean age = 16.7; Standard Deviation= ±1.4).

### 2.1. Procedures

Participants were required to sit in front of a laptop equipped with a microphone and asked to complete the following tasks devoted to accurately recording the participants' voice. Audio recording via computer were collected in which the subjects have been asked to read out loud a brief Aesop's fable, named "The northern wind and the sun" (hereafter referred to as the "tale task") from the laptop monitor and then to talk about how they spent the past week, or to recount any event they consider relevant, and to report for a minimum of 2 minutes (hereafter referred to as the 'Diary Task'). The reason this fable was chosen is that it is complete from a phonetic point of view, it is in fact widely used in phonetic descriptions of languages as an illustration of the spoken language. In the Handbook of the International Phonetic Alphabet and the Journal of the International Phonetic Alphabet, there is a translation of the fable in each language described, transcribed in the International Phonetic Alphabet.

The present work will focus on Whisper transcription of the "tale task"; the motivation behind this choice lies in the fact that in order to test Whisper performances it was necessary to start from a ground truth, i.e., the text of the fable, to be compared with Whisper final transcriptions. Data obtained through the Diary task have been analyzed and discussed in other works focused on Verbal Behavior Analysis that have been submitted and awaiting to be published.

## 3. Results

Once obtained Whisper transcription of the tale task for each participant, were calculated:

- Number of correct words (COR)
- Number of deleted words (DEL)
- Number of substituted words (SUB)
- Number of inserted words (INS)
- Word Error Rate (WER)

Figure 1 shows an example of elaborated text with highlighted errors of deletion, substitution, and insertion of words, while table 1 shows mean for each of the above-mentioned variables.

Figures 2, 3 and 4 represents the percentage of Correct, Deleted, Substituted, and Inserted words for each language (English, Italian and Russian). Percentages were calculated considering for each language the number of words composing Aesop's fable (119 for English, 116 for Italians and 96 for Russian).

**The North Wind and The Sun**

The north wind and the sun were *the* ***disputing (spitting)*** which was **the** stronger when ***a (the)*** ***traveller (triangle)*** came along wrapped in a warm *cloak* cloak they agreed that the one who first succeeded in making the ***traveller (triangle)*** take his cloak off should be considered stronger than the other then the north wind blew as hard as he could but the more he blew the more closely *he* did the ***traveller (triangle)*** ***fold (throw)*** his cloak around him and at last the north wind gave up the attempt then the sun shined **out** warmly and immediately the ***traveller (triangle)*** took off his cloak and so the north wind was obliged to confess that the sun was **the** stronger of the two.

**Figure 1:** Example of Whisper's errors. Words between * are Insertions, words between ** are Deletions, and words between *** are Substitutions. For instance, ***fold (throw)*** indicates that the word fold has been identified as a throw.

**Table 1**
Means of correct, deleted, substituted, and inserted words, identified by Whisper and the Word Error Rate. Means are shown for each participants' country (English, Italians and Russian) and split between healthy subjects (controls) and depressed patients (clinical groups).

| | Correct Words | | Deleted Words | | Substituted Words | | Inserted Words | | Word Error Rate | |
|---|---|---|---|---|---|---|---|---|---|---|
| | Mean | S. D. | Mean | S. D. | Mean | S. D. | Mean | S. D. | Mean | S. D. |
| **English controls** | 115.53 | 3.43 | 1.17 | 1.37 | 2.3 | 2.53 | 1.03 | 1.71 | 0.04 | 0.04 |
| **English clinical group** | 111.09 | 7.94 | 2.94 | 3.98 | 4.97 | 6.01 | 1.26 | 1.87 | 0.08 | 0.07 |
| **Italian controls** | 108.35 | 4.52 | 0.98 | 1.48 | 5.67 | 3.58 | 2.35 | 2.57 | 0.08 | 0.05 |
| **Italian clinical group** | 103.86 | 13.93 | 3.43 | 13.27 | 7.7 | 4.88 | 3.07 | 4.92 | 0.12 | 0.12 |
| **Russian controls** | 91.94 | 3.85 | 0.59 | 0.78 | 3.47 | 3.45 | 3.79 | 2.65 | 0.08 | 0.06 |
| **Russian clinical group** | 92.18 | 3.04 | 0.53 | 0.75 | 3.29 | 2.68 | 3.65 | 2.04 | 0.08 | 0.04 |

## 4. Discussion and Conclusion

In this work are reported descriptive analysis showing performances of the Automatic Speech Recognition System "Whisper" (OpenAI), used to transcribe data collected in different countries (United Kingdom, Italy, and Russia), in the context of a research protocol aimed at identifying features of speech expressions that may signal the presence of a depressive state, therefore speech data came (for each country) from controls and clinical groups of depressed subjects.

For each different language were calculate the Number of correct words (COR), Number of deleted words (DEL), Number of substituted words (SUB), Number of inserted words (INS) and the Word Error Rate (WER). The reason why we focused on these variables is related to the fact that speech recognition systems' performances are mostly susceptible to three types of errors, related to failures in discrete speech recognition, in continuous speech recognition, and in word spotting [1]. Errors in discrete speech recognition include deletion errors, when for instance the system ignores a word due to the speaker's failure in pronouncing it loudly and clearly enough, insertion errors when the system perceives noise as a speech unit, and substitution errors when the system identifies incorrectly a word [5]. Either for the English, Italian and Russian languages, were observed to have a high average of transcribed correct words, which consequently reflects lower errors in terms of deleted, substituted and inserted words; needs to be specified that the differences among the three languages in terms of correct

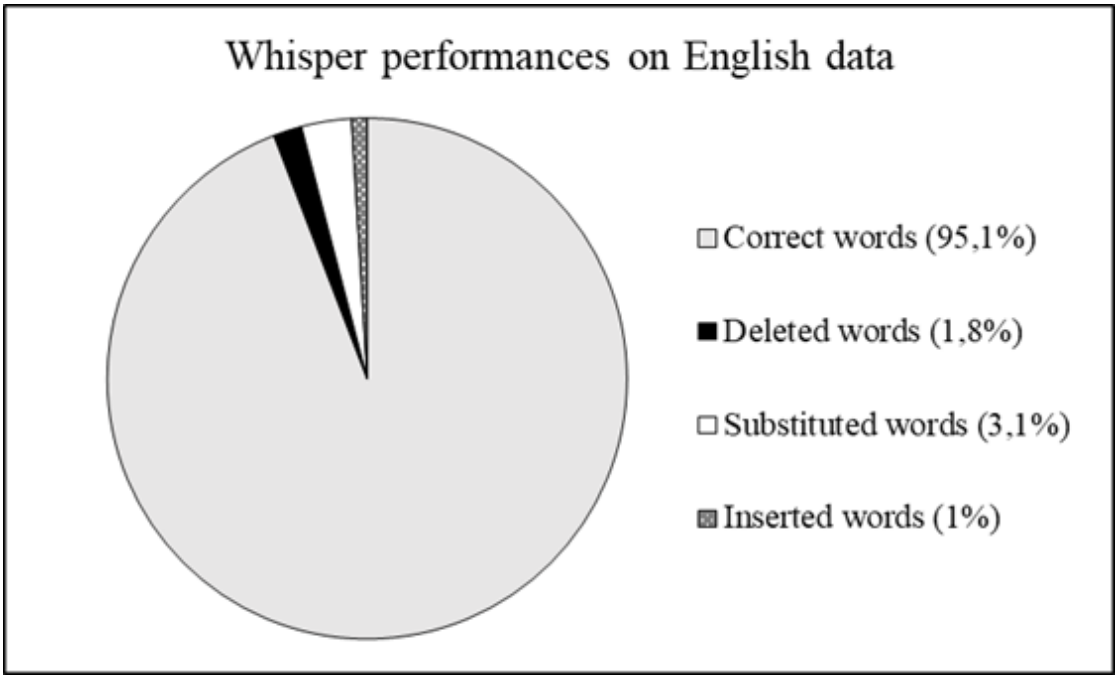


**Figure 2:** Percentages of Whisper's Correct, Deleted, Substituted, and Inserted words on English data.

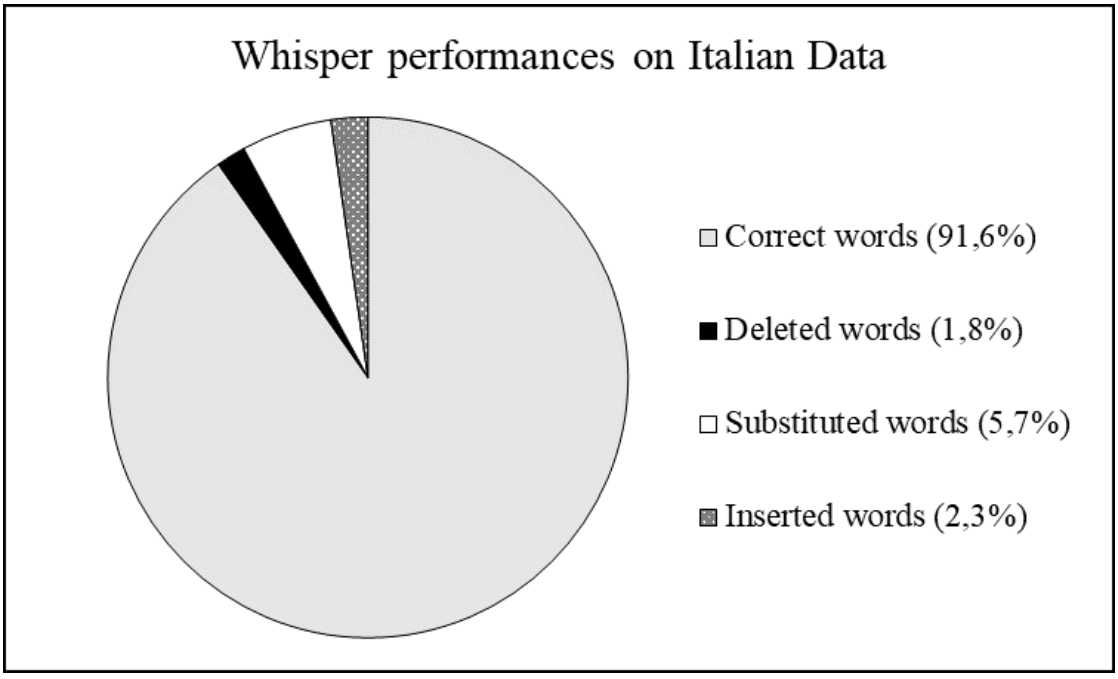


**Figure 3:** : Percentages of Whisper's Correct, Deleted, Substituted, and Inserted words on Italian data.


**Figure 4:** Percentages of Whisper's Correct, Deleted, Substituted, and Inserted words on Russian data.

words identified (showed in Table 1) are also due to the fact that for each language the number of words composing Aesop's fable is different (119 for English, 116 for Italians and 96 for Russian). Interestingly, it was observed that Whisper had higher performances when transcribing speech from the control groups of healthy subjects rather than participants leaving with depression, thus this happened with English and Italian languages while this effect was not observed with Russian language. For what concern differences between healthy and depressed users, within the English and Italian groups, it was observed that the control group tended to pronounce a higher number of correct words while reading the fable compared to depressed participants, while the latter tended to make more errors of word deletion and substitution compared to healthy subjects. However, this pattern was not observed with Russian participants; in fact, in this case, the clinical group pronounced a slightly higher number of correct words while reading the fable compared to healthy participants, and no huge differences were observed between healthy and depressed participants In terms of word deletion, substitution and insertion. For each language, and for each group, there were observed truly low Word Error Rates (WER).

The performance of a speech recognition system can be measured in terms of accuracy and speed. Accuracy, or performance accuracy, is known as word error rate (WER), that is a common metric of the speech recognition performance [1]. Our results concerning the WER are fairly promising since other studies [8] systematically comparing ASRs for clinical conversational speech found a wide range of word errors across the ASR engines, with values ranging from 35% to 65%; hopefully also our results will be helpful for clinical speech recognition, considering that automatic speech recognition can be a useful instrument to increase clinical documentation and clinician interventions, since transcription of audio recordings in psychotherapy would improve

therapy effectiveness, clinician training, and safety monitoring [9]. When observing Whisper performances for each language in terms of percentages of correct, deleted, substituted, and inserted words, without considering differences between controls and clinical groups, emerged that on the whole, Whisper showed high transcription performances (for each language the percentage of correct words transcribed is higher than 90%). For the English and Italian languages, the most committed mistakes were words substitutions, while for Russian there were observed higher percentages of both words substitution and insertion errors.

To conclude, testing the effectiveness of automatic speech recognition systems is a fundamental step in order to improve speech recognition technology and exploit it to support the enhancement of human interaction with machines. Fundamental is considering the importance of tailoring interfaces to users' needs and considering their nationality adopting a user-centered approach which can significantly enhance the user experience and increase the overall usability of computer systems and software. The aim is to create interfaces that are more welcoming, supportive, and effective in mitigating differences and fostering cooperation and collaboration within diverse groups of users. As a conclusive comment, we would like to highlight that in the present study, we have not specified how different levels of depressive disorders could have impacted the results of the task, since that was not the focus of the work, however, as already mentioned, we are also working on the results obtained from the diary task with the verbal behavior analysis, and in that case starting from DASS-21 (a self-report questionnaire aimed at measuring the level of depression, anxiety, and stress) [7] scores we considered participants with scores referring to moderate/severe/extremely severe levels of Depression as belonging to the clinical group, while patients with mild depression were excluded.

## Acknowledgments

The presented research has received funding from the following projects: EMPATHIC, EU H2020, N.769872; MENHIR, EU H2020, N. 823907; SIROBOTICS, MIUR, PNR 2015-2020, D.D. 1735/2017; ANDROIDS, Università della Campania "Luigi Vanvitelli" - V:ALERE 2019, D.R. 906/2019; SALICE, Università della Campania "Luigi Vanvitelli" - Giovani Ricercatori, D.R. 834/2022.

## References

[1] Shipra J. Arora & Rishi P. Singh, 2012. Automatic speech recognition: a review. International Journal of Computer Applications 60.9

[2] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, Michael Auli, 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. Advances in neural information processing systems, 33, 12449-12460.

[3] Aaron T. Beck, Robert A. Steer, and Gregory Brown, 1996. Beck depression inventory–II. Psychological assessment.

[4] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Jian Wu, Michael Zeng, Xiangzhan Yu, Furu Wei, 2022. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. IEEE Journal of Selected Topics in Signal Processing, 16(6), 1505-1518.

[5] John Levis & Ruslan Suvorov, 2012. Automatic speech recognition. The encyclopedia of applied linguistics.

[6] Jennifer C. Lai, Marie Karat, Nicole Yankelovich, 2009. Conversational speech interfaces and technologies. Human-Computer Interaction. CRC Press, 71-82.

[7] P.F. Lovibond & S.H. Lovibond, 1995. The structure of negative emotional states: Comparison of the Depression Anxiety Stress Scales (DASS) with the Beck Depression and Anxiety Inventories. Behaviour research and therapy, 33(3), 335-343.

[8] Jodi Kodish-Wachs, Emin Agassi, Patrick Kenny, J. Marc Overhage, 2018. A systematic comparison of contemporary automatic speech recognition engines for conversational clinical speech. In AMIA Annual Symposium Proceedings (Vol. 2018, p. 683). American Medical Informatics Association.

[9] Adam S. Miner, Albert Haque, Jason A. Fries, Scott L. Fleming, Denise E. Wilfley, G. Terence Wilson, Arnold Milstein, Dan Jurafsky, Bruce A. Arnow, W. Stewart Agras, Li Fei-Fei, Nigam H. Shah, 2020). Assessing the accuracy of automatic speech recognition for psychotherapy. NPJ digital medicine, 3(1), 82.

[10] Alec Radford, Jong W. Kim, Tao Xu, Greg Brockman, Christine McLeavey, Ilya Sutskever, 2022. Robust speech recognition via large-scale weak supervision. arXiv preprint arXiv:2212.04356.

[11] Robert D. Rodman, 1999). Computer speech technology. Norwood, MA: Artech House.

[12] Yla R. Tausczik, and James W. Pennebaker, 2010. The psychological meaning of words: LIWC and computerized text analysis methods. Journal of language and social psychology, 29(1), 24-54.

[13] Jeffrey E. Young, Janet S. Klosko, and Marjorie E. Weishaar, 2006. Schema therapy: A practitioner's guide. guilford press.