FONTI 4.0: Evaluating Speech-to-Text Automatic Transcription of Digitized Historical Oral Sources

Roberta Bianca Luzietti[1], **Niccolò Pretto**[1], **Frédéric Kaplan**[2], **Alain Dufaux**[2] **and Sergio Canazza**[1]
1. University of Padua, Italy
2. École Polytechnique Fédérale de Lausanne
robertabianca.luzietti@unipd.it,
{niccolo.pretto,canazza}@dei.unipd.it,
{frederic.kaplan,alain.dufaux}@epfl.ch

Abstract

Conducting "manual" transcriptions and analyses is unsustainable for most historical oral archives because they require a remarkable amount of funds and time. The FONTI 4.0 project aims at exploring the suitability of automatic transcription and information extraction technologies for making historical oral sources available. In this work, we conducted an experiment to test the performance of two commercial speech-to-text services (Google Cloud Speech-to-text and Amazon Transcribe) on digitized oral sources. We created an eight-hour corpus made of manually transcribed and annotated historical speech recordings in TEI format. The results clearly show how audio quality and disturbing elements (e.g., overlaps, foreign words, etc.) impact on the automatic transcription, showing what needs to be improved for implementing an unsupervised transcription chain.

1 Introduction

FONTI 4.0[1] is a project aiming at exploring the suitability of automatic transcription and analysis tools for the preservation of historical oral sources recorded on analog carriers, in particular magnetic tapes. The digitization of an audio archive is a long and expensive task that can require several years. Furthermore, the content of audio recordings needs to be listened and cataloged for making audio recordings retrievable. Archives composed by hundreds or thousands of hours of audio require a huge amount of time, people and funds for making the content accessible and preventing their exploitation. Therefore, automatizing the transcription and the analysis task could drastically reduce the time for making digitized audio recordings accessible.

The project consists in a transcription-chain (T-chain), firstly defined in (van Hessen et al., 2020), that differs in two main aspects: (a) in FONTI 4.0, the transcription obtained with speech-to-text (STT) algorithms should not be corrected by human; (b) an additional restoration step could be required for digitized audio recordings. Furthermore, differently from STT evaluation experiments conducted by (Moore et al., 2019; Kostuchenko et al., 2019; Filippidou and Moussiades, 2020), we decided to employ two commercial software, namely Google Cloud Platform and Amazon Web Services, to test their ability to transcribe historical analog recordings, and to eventually include in our pipeline.

During the digitization process, speed and equalization errors can occur, especially when different speed and equalization configurations are used in different part of the same tape (Pretto et al., 2020). This leads to distortions of the recorded signal that becomes unlistenable. By using the correction workflow and digital filters described in (Pretto et al., 2021a; Pretto et al., 2021b) these errors can be corrected and at least parts of the signal can be saved. This task is essential for making the speech signal suitable for STT algorithms. This paper aims at evaluating the transcription performance of two commercial software on a real use case and identifying potential problems or limitations concerning peculiarities of analog audio recordings. Section 2 describes the corpus, used

Copyright © 2021 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

[1] csc.dei.unipd.it/fonti40en/ (last accessed September 2nd, 2021)

CEUR-WS.org/Vol-3033/paper45.pdf

as ground-truth for the experiment. Section 3 outlines the methodology adopted for this experiment, whereas results are reported in Section 4. Finally, Section 5 presents the authors' conclusions.

2 Corpus

The *Cinema & Civiltà* (C&C) corpus was conceived within the FONTI 4.0 as ground-truth for evaluating the performance of STT services on a real case study. To build the corpus, we transcribed speech recorded on four magnetic tapes made available by the Giorgio Cini Foundation in Venice and digitized at the Centro di Sonologia Computazionale - CSC (Canazza and De Poli, 2020). The recordings are parts of the *Cinema & Civiltà* conference for the awarding of the San Giorgio prize, part of the Venice Film Festival, that took place between the 7^{th} and 9^{th} of September 1959, attended by important figures of the history of cinema such as Roberto Rossellini and representatives of the Italian literary critics such as Vittore Branca. Each reel of magnetic tape is composed of two sides: each side counting 60 minutes of recorded speech for a total of eight hours of recording. The C&C corpus is also a multilingual corpus of 64,930 tokens and three subcorpora: Italian 49,772 tokens, French 9,555 tokens (L1 and L2), and Spanish 5,603 tokens. This corpus was manually transcribed and annotated as described in the following subsections and is available at this link[2].

2.1 Transcription

Defining the methodology for the transcription is an important step for the preservation, analysis and access of oral sources. The main difficulty consists in making decisions on how to represent and convey both verbal and non-verbal elements in written form. Because of the absence of a universal standard of transcription (Schorrsidt, 2011), the methodology usually depends on the research aim.

In this research, we decided to complete a verbatim transcription, by reporting every word spoken in the recording including errors, false starts, truncations, and overlaps in Italian, French and Spanish. Using the software ELAN (Lausberg and Sloetjes, 2016), we first segmented audio files extracted from the digitized tapes, making the start and end of each segment coincide with the speaker's turn of talk. Then, we transcribed each segment while listening to the corresponding part of audio in slow motion. Eventually, we opted for employing automatic transcriptions from Google Cloud Speech-to-text (GCS) and Amazon Transcribe (AT)[3], later used in the STT experiment, and correcting the text playing the audio at normal speed. This allowed us to save half the time for each transcription, which previously required a full day of work. Moreover, we were able to retrace and match the identity of the speakers to the voices in the recordings, through the consultation of historical documentation on the conference, and also by comparing voices across the recordings.

2.2 Annotation

The annotation was employed for the addition of important metadata to the C&C corpus regarding different levels of audio quality and the presence of disturbing elements in the recordings. Our methodology is in compliance with the Text Encoding Initiative (TEI) standard guidelines[4] for transcribed spoken material (Burnard and Bauman, 2007). To proceed with the annotation, we first converted the transcription files from the ELAN .eaf into the XML TEI standard using the EXMARaLDA (Schmidt and Wörner, 2014) tool TEI Drop (Schorrsidt, 2011). Subsequently, we used Oxygen[5] to assign TEI tags to the relevant tokens. The list of tags together with a brief description and examples is given below:

<**pause**> marks a pause either between or within utterances in the same segment, **e.g.**: *unica fisionomia. <pause/> Parte dell'architettura*;

<**unclear**> contains a word, phrase, or passage that could not be transcribed with certainty because it is illegible or inaudible in the source, **e.g.**: *gli stessi <unclear reason="inaudible"> strumenti </unclear>, volti agli stessi fini*;

<**gap**> indicates a point where material has been omitted in the transcription because it is inaudible, **e.g.**: *erba che sorgerà <gap reason="inaudible"/> quell'asfalto.*;

[2]DOI: 10.5281/zenodo.5645827

[3]Automatic transcriptions were obtained on the 16th, 17th, 19th and 24th of March 2021.

[4]tei-c.org/release/doc/tei-p5-doc/en/html/TS.html (last accessed September 3rd, 2021

[5]oxygenxml.com (last accessed September 3rd, 2021)

<**foreign**> identifies a word or phrase as belonging to some language other than that of the surrounding text, **e.g.**: *<foreign xml:lang="fr-FR"> Mesdames, messieurs </foreign>*;

<**shift**> marks the point at which some paralinguistic feature of a series of utterances by any one speaker changes, **e.g.**: *Io credo che questo argomento sia <shift feature="tempo" new="a"/> particolarmente importante <shift feature="tempo" new="normal"/> per vedere*;

<**del**> contains a letter, word, or passage indicated as superfluous by the annotator, in this case it was used for false starts, repetitions and truncations, **e.g.**: *in questo <del type="falseStart"> moden </del> momento* (false start) *momento di <del type="repetition"> di </del> crisi* (repetition) *suggestione di <del type="truncation"> spettaco </del> di spettacolo (truncation)*;

<**anchor**> was used to mark overlaps by attaching an identifier to a point within a text, **e.g.**: *a contatto di un <anchor synch="ovrl6" xml:id="S06"/> pensiero <anchor synch="ovrl6e" xml:id="S06e"/> lo inducono a (interrupted speaker) <anchor xml:id="ovrl6"/> Io non lo vedo. Chi è questo? Chi è questo? <anchor xml:id="ovrl6e"/> (interrupting speaker)*;

<**distinct**> identifies any word or phrase which is regarded as linguistically distinct, as in the case of prosodically unified units, **e.g.**: *staccarsi da <distinct type="pcu"> questa estetica </distinct> e dai pregiudizi*;

<**vocal**> marks any vocalized but not necessarily lexical phenomenon, **e.g.**: *del nostro mondo <vocal> <desc>cough</desc> </vocal> che direi postmoderno.*;

<**incident**> marks any phenomenon or occurrence, not necessarily communicative, for example incidental noises or other events affecting communication, **e.g.**: *è attività creatrice, <incident><desc>noise</desc></incident> ma non propriamente l'artista*;

<**note**> contains notes or citations, and, for the purpose of this research, it was used to annotate the audio quality at the beginning of each segment, **e.g.**: *<note>good </note>*;

Audio quality annotations (<note>) were assigned to each segment using the the following scale (Samar and Metz, 1988):

excellent: speech is completely intelligible;

good: speech is intelligible with the exception of a few words or phrases;

fair: with difficulty, the listener can understand about half the content of the message;

poor: speech is very difficult to understand, only isolated words or phrases are intelligible;

bad: speech is completely unintelligible.

The distribution of words (without punctuation and events) for each audio quality annotation is reported in Table 1.

Scale	it-IT	fr-FR	es-ES	TOT
Excel.	9,075	5,930	4,097	19.102
Good	30,571	2,514	800	33.885
Fair	2,919	83	0	3,002
Poor	1,417	23	0	1,440
TOT	43,984	8,550	4,897	0

Table 1: Number of words (no punctuation nor events) annotated with different audio quality tags.

3 Experiments

The STT experiment consisted in testing the ability of GCS and AT to correctly transcribe historical recordings. Furthermore, we decided to investigate the performance of STT transcriptions obtained from GCS and AT at different levels of audio quality and in presence of disturbing elements in the recordings such as background noise, overlaps, code switching etc. (see Section 2.2).

To analyze the performance of the two STT systems, we developed a Jupyter notebook able to filter the text by language, audio quality, disturbing elements, etc., and select several options, such as tokenization rules. In this experiment, we decided to use only lower case characters, split apostrophes and remove punctuation from both manual and automatic transcriptions. The ground-truth and the resulting transcription of the STT

services were canonicalized. The alignment algorithm works on single utterances and minimizes the Levenshtein distance (Jurafsky and Martin, 2008). The obtained metrics were: the number of correct matches (COR) and mismatches, i.e.: deletions (DEL), substitutions (SUB) and insertions (INS), and the word error rate (WER), which is the ratio between the number of mismatches and words in the reference text (Morris et al., 2004). It is important to note that we did not employ this metric to tell how good a system is, but only that one is better than the other (Errattahi et al., 2018).

In order to avoid the introduction of errors not due to the transcription task, we decided not to use the automatic language recognition feature because it could drastically impact on the performance. Therefore, we cut and divided the audio files in different languages and automatically transcribed them separately.

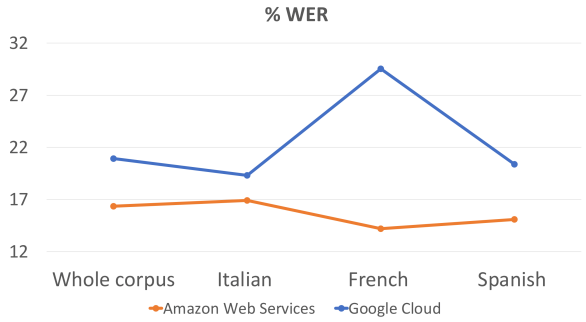

Figure 1: WER of GCS and AT transcriptions on the whole corpus and sub-corpora.

STT	WER	COR	DEL	SUB	INS
AT	16.35%	49,480	2,639	**5,312**	**1,440**
GCS	20.92%	46,510	**5,837**	5,084	1,094

Table 2: Word error rate (WER), Correct matches (COR), deletions (DEL), substitutions (SUB) and insertions (INS) of the Amazon Transcribe (AT) and Google Cloud Speech-to-text (GCS) transcriptions of the overall C&C corpus.

4 Results

In this preliminary work we illustrate and compare mainly WER trends between the two STT systems, calculated on the entire corpus as well as each sub-corpora in relation to audio quality levels and the presence of disturbing elements.

Figure 1 illustrates that the performance of AT are better than GCS in all corpora. The difference between the two systems is small in the Italian sub-corpus, but much wider in the French. A possible explanation could be the presence of L2 speakers of French whose pronunciation could have negatively affected the recognition performance. Nevertheless, it should be also considered that the Italian sub-corpus is more than five times bigger than the French and the Spanish.

STT software performance can be further observed in Table 2: for the transcription of the whole corpus, AT scores a lower WER and finds more correct matches than GCS. On the other hand, deletions in GCS are more than double than in AT, whereas substitutions and insertions are higher in AT than in GCS. In any case, the number of deletions and insertions between AT and GCS are different probably because the two services make use of different language model weights.

Figure 2 shows that transcription performance are very similar in Italian and Spanish with "Excellent" quality, but not in French. For this reason, we cannot impute the bad GCS performance to audio quality. In the Italian sub-corpus, performance are also similar with "Good" quality, but not in the Spanish, where both services performed badly. The negative impact of audio quality is also evident in the French sub-corpus, despite WER values are much higher than Italian.

Results in Figure 3 display the annotated disturbing events found in the C&C corpus that were assumed to negatively affect the performance of STT software in terms of WER. The element that provides the minor disturbance is shift, although the scored WER value for this tag is higher than the one calculated on the overall evaluation. About the other disturbing elements, they show a major impact on the transcription of both STT services. Overall, AT performance is better with most disturbing elements. The only exception is represented by code-switching events in foreign languages for which GCS had a better performance.

5 Conclusion

In this article we conducted a preliminary research experiment testing the ability of STT software to correctly transcribe digitized historical oral sources on magnetic tape. It should be noted, that since this preliminary work has been conducted on a small sample of data, our results are

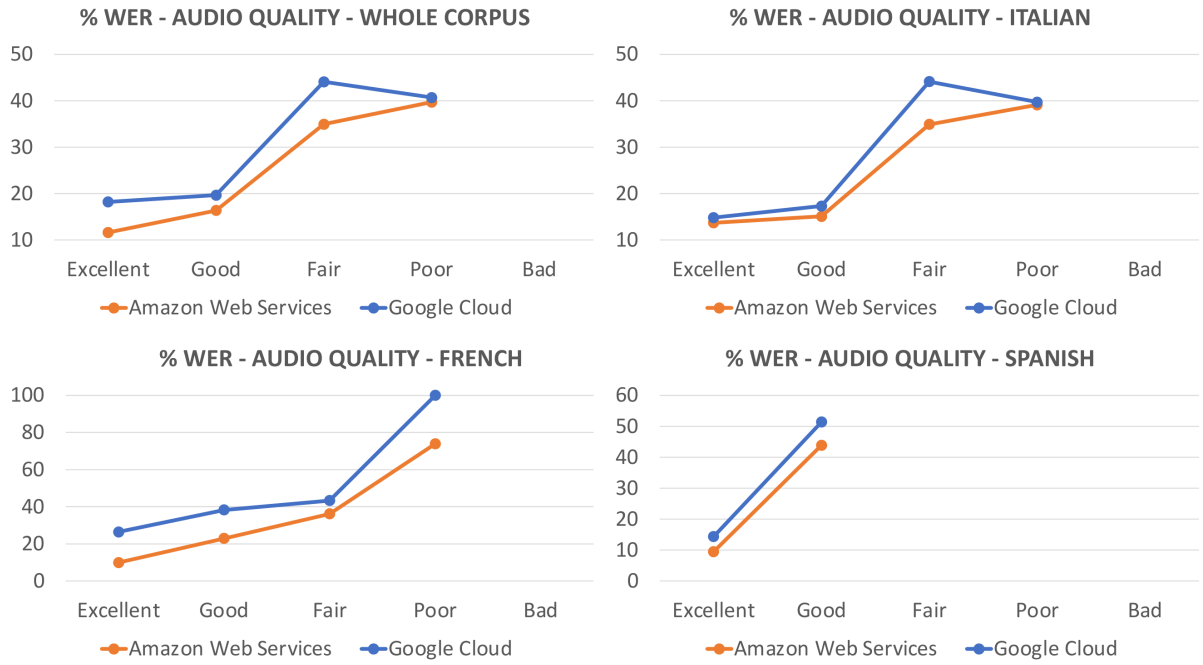

Figure 2: WER of GCS and AT with different audio quality - whole corpus and sub-corpora.

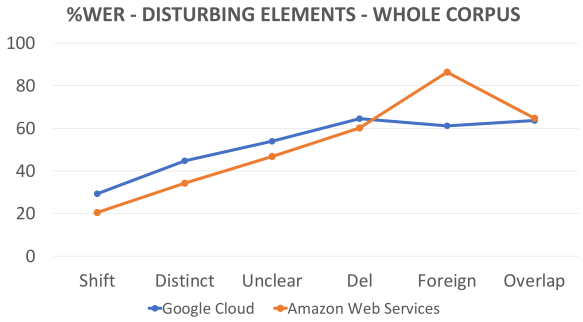

Figure 3: WER of GCS and AT with elements on the whole corpus and sub-corpora.

only indicative of which elements represent the biggest obstacle for STT software performance.

In spite of disturbing elements and the variation of audio quality in the recordings, we demonstrate that with our dataset and in terms of WER, AT performed more accurate transcriptions compared to GCS . On the other hand, GCS was better at recognizing foreign words. Table 2 shows that AT introduces less incorrect words but more insertions and substitutions. This should be taken into consideration when working with automatic information extraction tools (e.g., Named Entity Recognition algorithms) applied to automatic transcriptions. Further analysis should investigate the cause of this trend, to verify if this behavior is also due to alignment or tokenization errors.

With respect to software performance evaluations in relation to variables characterizing analog recordings of speech, we found evidence that audio quality drastically impacts on the number of mismatches. Observations about the incidence of disturbing elements, on the other hand, cannot be generalized since sub-corpora are in three different languages and have three different sizes. Throughout the analyses we noted that the most negative impact on transcription, in terms of the increase of WER, is caused by the presence of some specific recurring elements, i.e.: code-switching (foreign), overlaps and probably even the production of L2 speakers (Figure 3). Nonetheless, given the necessity of preserving historical documents in a more time and cost effective way, we came to the conclusion that researchers working on the preservation of historical recordings will benefit from the use of the T-chain. This is because the reduction by half of the time required for manual transcriptions in slow motion does compensate the lack of accuracy. This means that researchers working on the collection and preservation of oral archives will be able to focus on filling the gap between human and machine output.

Further contributions will be necessary for conducting experiments on L1 and L2 data separately, cross-language testings reducing the Italian subset to the size of the French and Spanish sub-corpora

and evaluating the impact of incorrect transcriptions on WER. Language identification through code-switching is another important problem for automatic transcription. Both services recently provided this functionality, but while we are writing this paper, the Google Cloud is still a preview version. As soon as the feature will be available the performance of automatic language recognition algorithms should also be investigated, especially because this feature is essential for automatizing the transcription of entire archives.

Acknowledgments

This paper is produced under the FONTI 4.0 project, financed by resources from the Regional Operational Program co-financed with the European Social Fund 2014-2020 of the Veneto Region. An important acknowledge is due to Fondazione Giorgio Cini, Venice for making available its precious audio material as well as for its help in the recording analysis, and Matteo Pettenò for his contribution in the Jupyter notebook development.

References

Lou Burnard and Syd Bauman, editors, 2007. *TEI P5: Guidelines for Electronic Text Encoding and Interchange*, chapter A Gentle Introduction to XML. Text Encoding Initiative Consortium.

Sergio Canazza and Giovanni De Poli. 2020. Four Decades of Music Research, Creation, and Education at Padua's Centro di Sonologia Computazionale. *Computer Music Journal*, 43(4):58–80, 10.

Rahhal Errattahi, Asmaa El Hannani, and Hassan Ouahmane. 2018. Automatic speech recognition errors detection and correction: A review. *Procedia Computer Science*, 128:32–37.

Foteini Filippidou and Lefteris Moussiades. 2020. A benchmarking of ibm, google and wit automatic speech recognition systems. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pages 73–82. Springer.

Daniel Jurafsky and James Martin. 2008. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*, volume 2. Pearson, Prentice Hall, 02.

Evgeny Kostuchenko, Dariya Novokhrestova, Marina Tirskaya, Alexander Shelupanov, Mikhail Nemirovich-Danchenko, Evgeny Choynzonov, and Lidiya Balatskaya. 2019. The evaluation process automation of phrase and word intelligibility using speech recognition systems. In *International Conference on Speech and Computer*, pages 237–246. Springer.

Hedda Lausberg and Han Sloetjes. 2016. The revised neuroges–elan system: An objective and reliable interdisciplinary analysis tool for nonverbal behavior and gesture. *Behavior research methods*, 48(3):973–993.

Meredith Moore, Michael Saxon, Hemanth Venkateswara, Visar Berisha, and Sethuraman Panchanathan. 2019. Say what? A dataset for exploring the error patterns that two ASR engines make. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2019-Septe:2528–2532.

Andrew Cameron Morris, Viktoria Maier, and Phil Green. 2004. From wer and ril to mer and wil: improved evaluation measures for connected speech recognition. In *Eighth International Conference on Spoken Language Processing*.

Niccolò Pretto, Alessandro Russo, Federica Bressan, Valentina Burini, Antonio Rodà, and Sergio Canazza. 2020. Active preservation of analogue audio documents: A summary of the last seven years of digitization at csc. In *Proceedings of the 17th Sound and Music Computing Conference, SMC20*, pages 394–398, Torino, Italy.

Niccolò Pretto, Nadir Dalla Pozza, Alberto Padoan, Anthony Chmiel, Kurt James Werner, Alessandra Micalizzi, Emery Schubert, Antonio Rodà, Simone Milani, and Sergio Canazza. 2021a. A workflow and novel digital filters for compensating speed and equalization errors on digitized audio open-reel tapes. In *Proceedings of Audio Mostly 2021, AM21*, Trento, Italy.

Niccolò Pretto, Edoardo Micheloni, Anthony Chmiel, Nadir Dalla Pozza, Dario Marinello, Emery Schubert, and Sergio Canazza. 2021b. Multimedia Archives: New Digital Filters to Correct Equalization Errors on Digitized Audio Tapes. *Advances in Multimedia*, 2021:5410218.

Vincent J Samar and Dale Evan Metz. 1988. Criterion validity of speech intelligibility rating-scale procedures for the hearing-impaired population. *Journal of Speech, Language, and Hearing Research*, 31(3):307–316.

Thomas Schmidt and Kai Wörner. 2014. Exmaralda. In Jacques Durand, Ulrike Gut, and Gjert Kristoffersen, editors, *The Oxford handbook of corpus phonology*. Oxford University Press.

Thomas Schorrsidt. 2011. A tei-based approach to standardising spoken language transcription. *Journal of the Text Encoding Initiative*, 1.

Arjan van Hessen, Silvia Calamai, Henk van den Heuvel, Stefania Scagliola, Norah Karrouche, Jeannine Beeken, Louise Corti, and Christoph Draxler.

2020. Speech, voice, text, and meaning: A multidisciplinary approach to interview data through the use of digital tools. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*, ICMI '20 Companion, page 454–455, New York, NY, USA. Association for Computing Machinery.