

Bringing NURC/SP to Digital Life: the Role of Open-source Automatic Speech Recognition Models

Lucas Rafael Stefanel Gris¹, Arnaldo Candido Junior²,
Vinícius G. dos Santos³, Bruno A. Papa Dias³, Marli Quadros Leite³,
Flaviane Romani Fernandes Svartman³, Sandra Aluísio³

¹Federal University of Goiás, Brazil

²São Paulo State University, Brazil

³University of São Paulo, Brazil

lucas.gris(at)ufg.discente.br, arnaldo.candido(at)unesp.br
{brunoadiaspapa1, mqleite, flavianesvartman}(at)usp.br
sandra(at)icmc.usp.br

Abstract. *The NURC Project that started in 1969 to study the cultured linguistic urban norm spoken in five Brazilian capitals, was responsible for compiling a large corpus for each capital. The digitized NURC/SP comprises 375 inquiries in 334 hours of recordings taken in São Paulo capital. Although 47 inquiries have transcripts, there was no alignment between the audio-transcription, and 328 inquiries were not transcribed. This article presents an evaluation and error analysis of three automatic speech recognition models trained with spontaneous speech in Portuguese and one model trained with prepared speech. The evaluation allowed us to choose the best model, using WER and CER metrics, in a manually aligned sample of NURC/SP, to automatically transcribe 284 hours.*

Index Terms: NURC/SP corpus, automatic speech recognition evaluation, Portuguese language, spontaneous speech

1. Introduction

The Cultured Urban Linguistic Norm (NURC) Project (*Norma Linguística Urbana Culta*, in Portuguese) [Oliveira Jr. 2019] was conceived in 1969 with the objective of studying the cultured urban norm spoken in five Brazilian capitals — Recife¹, Salvador, Rio de Janeiro, São Paulo², and Porto Alegre. It was responsible for the compilation of a large corpus for each capital, around three hundred hours, and with a varied recording quality, taking inquiries of three different textual genres. The informers have an academic degree, are born in the city under study, and are children of native Portuguese speakers. There is an equilibrium between the male and the female gender and there are three informer age groups: from 25 to 35 years old, from 36 to 55 years old, and more than 56 years old.

The NURC/SP, the focus of this article, was organized in three phases, the first was executed between 1972 and 1982 and was responsible for the selection of the speakers, transcription of several recordings and first analyses. The phase 2 lasted from 1983 to 2021 and produced 14 books that deal with the description of the phonetic, phonological, grammatical, and lexical aspects of the cultured urban spoken language and orality

¹<https://fale.ufal.br/projeto/nurcdigital/>

²<https://nurc.fflch.usp.br/>

and teaching. From 2014 to 2017, the “*Alexandre Eulalio*” Center for Cultural Documentation (CEDAE/UNICAMP), digitized the collection, originally recorded in analog medium, and made part of the files available on the Web, in the Sound Documents Platform. The files digitized by CEDAE were given in October 2020 to the Tarsila project³, which revisited NURC/SP with the objective of making available a protocol with automatic speech processing tools to speed up the full public availability of this large corpus and thus leverage the development and study of Brazilian Portuguese (BP) speech processing methods. Therefore, we mark the start of phase 3 of NURC/SP in 2021. More specifically, once NURC/SP is fully processed, it will be possible to use the corpus as a training dataset to build automatic systems for spontaneous speech recognition, in addition to facilitating linguistic studies, given its future availability on a portal that will allow specific searches. The protocol that guides the NURC/SP processing is strongly based on the protocol developed by the NURC Digital project [Oliveira Jr. 2016] to process NURC-Recife, but it incorporates speech processing systems to speed up the processing of this large corpus. While the NURC-Recife corpus took seven years to be processed and publicly available, we hope that using speech processing tools will help to bring NURC/SP quickly to digital life. As examples of these systems, there are the automatic speech recognizer (ASR) that was chosen in this study — the model trained with the corpus CORAA ASR [Candido Junior et al. 2021], the aeneas forced aligner used to synchronize audio and transcription⁴, and the forced phonetic aligner Alinha-PB⁵ used in conjunction with an utterance segmentation method based on prosody.

The digitized NURC/SP is composed of 375 inquiries, totaling approximately 334 recording hours. Although 47 inquiries have transcription, there was no alignment between audio and transcription, needed for creating ASR models, and 328 inquiries were not transcribed, relying solely on audio. Given these characteristics, in the Tarsila project, NURC/SP was divided into three working subcorpora: (i) **the Minimum Corpus (MC)**, with 21 inquiries, extensively studied in phase 1, and whose audios and transcripts were automatically aligned by the forced aligner aeneas and manually segmented into prosodic utterance units by a team of six annotators, based on work of [Raso and Mello 2012]. The annotation of the MC is described in [Santos et al. 2022] and the corpus is publicly available on the Portulan Clarin repository under CC BY-NC-ND 4.0 license (<https://hdl.handle.net/21.11129/0000-000F-73CA-C>); (ii) **the Corpus with non-aligned Audios and Transcripts**, composed of 26 inquiries that will be automatically segmented and manually revised. This corpus and alignment method will be described in another article; and (iii) **the Audio Corpus**, composed of 328 audios that were not transcribed, the focus of this article.

The main contributions made in this work are summarised as follows:

1. to present the NURC/SP corpus, the challenges to its preprocessing using ASR tools aiming at its public release to enlarge the public availability of spontaneous speech for ASR training in BP;
2. to evaluate, using WER (Word Error Rate) and CER (Character Error Rate) metrics, four public available BP ASR models in a manually aligned sample of NURC/SP; and

³<https://sites.google.com/view/tarsila-c4ai>

⁴<https://www.readbeyond.it/aeneas/>

⁵<https://conversoralinhador.herokuapp.com/>

3. to make publicly available the code used for the evaluation and results at <https://github.com/nilc-nlp/nurc-sp>.

The four open-source ASR models are described in Section 2.2.3 and were assessed in this article for the selection of the best (Section 3.2) to automatically transcribe 328 inquiries of NURC/SP, totaling 284 hours. The assessment was performed on a representative sample of the NURC/SP Minimum Corpus, which is described in Section 2.1. This article also presents the challenges related to the text genres and audio quality, bringing an error analysis (Section 3.3), based on the transcription rules of the NURC project [Preti 1999], in all the inquiries of the Minimum Corpus of NURC/SP. Finally, in Section 4, we present the conclusions and future work.

2. Background

2.1. The Minimum Corpus of NURC/SP

The three textual genres used in MC and in NURC in general are: formal elocutions (EF), for instance, lectures and conference talks; informal dialogues between the speakers, with the presence of a documenter (D2); and interviews about different topics, carried out by an interviewer with the interviewee (DID). They bring an interesting phenomenon of spontaneous speech which is the degree of overlap in the spectrum (little - average - much overlap), generally correlated with the textual genre. EF-type inquiries have little overlap (of accidental speakers, such as students), DID has a medium overlap, as it involves two speakers, one of them being the interviewer, and D2 is likely to have a lot of overlapping, as there are two speakers, in addition to the mediator.

Table 1 presents the characteristics of the 21 MC inquiries, separated into the three textual genres (1st column), audio quality (2nd column) which is summarized in 3 classes (+, -, mixed), its duration (3rd column), and a short description of the audio quality (last column). Regarding audio quality, we merged two pieces of information: (1) the audio volume (good, very good, audible, low) and (2) the quality of the recording in relation to the voice of the speakers (clear, good, medium bass, shaky voice, distorted) and with regard to external events such as car noise, accidental speaker interference, hiss and background music. In Table 1 we have 10 inquiries classified as + (good, very good, clear, audible, clear), 6 inquiries as - (a little low, low, with distortion, shaky voice), and 5 inquiries as mixed, as they bring characteristics of the 2 previous classes.

2.2. ASRs Models

Automatic Speech Recognition (ASR) consists on generating text from audio signals [Karpagavalli and Chandra 2016]. ASR systems can be used in personal assistants, closed caption on television and streaming, customer attendance tools, voice dialing, structured document creation, among others [Li et al. 2015]. Deep learning led to many breakthroughs in ASR research, with noticeable advancements, specially after the use of self-supervised techniques, improving results for the task. A self-supervised state-of-the-art architecture is Wav2vec 2.0, created by Meta AI [Baeovski et al. 2020]. In self-supervised learning, the model should automatically organize the input structure during the pre-training. In the ASR context, this representations can be used to transcribe audios using less labelled data. In this work, we used Wav2vec 2.0 based models.

Table 1. Characterization of the 21 inquiries of the MC by text gender and audio quality. The five inquiries in bold were used to evaluate four ASR models trained with BP speech.

Text gender	Audio quality	Duration	Description related to both the voice of the speakers and external events
EF_388	+	01:01:10	very good audio
EF_153	+	01:11:11	very good audio
EF_156	+	1:35:37	very good sound
EF_124	+	01:35:37	clear audio
EF_405	Mixed	00:30:51	good and low audio parts; some interference from accidental speakers
EF_377	-	00:30:40	low audio
DID_242	+	00:44:08	clear audio
DID_234	+	00:36:22	clear audio
DID_235	+	00:34:49	clear audio
DID_137	+	00:38:44	clear, good audio
DID_018	Mixed	00:54:40	audible, with distortion in some recording snippets
DID_251	Mixed	00:36:39	clear audio; lower audio parts
DID_208	Mixed	00:44:51	good sound, a little low
DID_161	-	00:40:36	low audio
DID_250	-	00:40:34	very low audio
D2_333	+	00:56:07	clear audio
D2_255	+	1:24:01	clear sound
D2_396	Mixed	01:10:15	clear sound; car noises in the background
D2_062	-	1:23:42	the audio is not good as the voice is shaky
D2_360	-	1:03:32	a little bit low audio
D2_343	-	01:20:23	low audio

2.2.1. Wav2vec

Wav2vec 2.0 is an end-to-end architecture inspired in the work of [Schneider et al. 2019]. The architecture is presented on Figure 1. In general, the training process can be divided into self-supervised pre-training and supervised fine-tuning. Pre-training consists

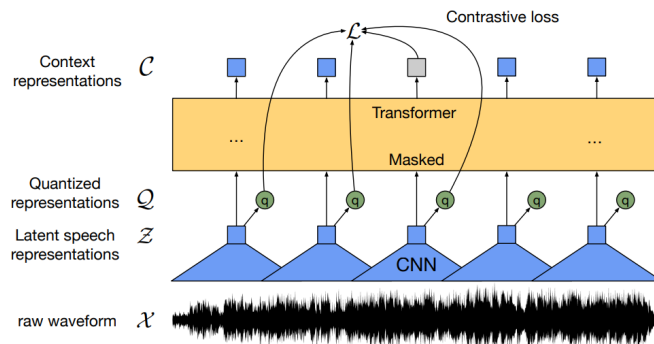


Figure 1. Wav2vec 2.0 Architecture (adapted from [Baevski et al. 2020]).

on masking the speech signal while the model should predict the masked parts. Wav2vec uses a contrastive method applied to differentiate between quantized input representations from a set of distractors. This results in discrete speech representations that can be fine-tuned to the ASR task. For the supervised fine-tuning, a projection layer is included

as last layer of the model including speech labelled at sentence level. The architecture allows to the discrete representations to be used as input for the Transformer attention mechanism [Vaswani et al. 2017]. The architecture presented in Figure 1 is divided as follows. A multi-layer convolutional encoder $f : X \mapsto Z$ maps the raw speech signal X in latent speech representation da fala z_1, \dots, z_T in T timesteps. The encoder output is then sent to the Transformer block, applying the mapping $g : Z \mapsto C$ that converts latent representations Z into contextualized representations c_1, \dots, c_T . Wave2vec 2.0 has different versions, for example the models BASE (92M parameters) and LARGE (317M parameters).

2.2.2. Corpora for Portuguese ASR

Several corpora for ASR in BP have been released in the last years. In this section, we focus on resources used to train the models we applied on NURC/SP raw audios.

One of this corpora is CETUC [Alencar and Alcaim 2008], a corpus with 145 recorded hours of speakers of both sexes, 50 males and 50 females. In total, the corpus contains 1.000 phonetic-balanced sentences, extract from the CETEN-Folha corpus. The audios are sampled at 16kHz and are public available⁶. Multilingual LibriSpeech (MLS) [Pratap et al. 2020] is a corpus containing Librivox audiobooks⁷. MLS is public available using the Creative Commons (CC) BY 4.0 and encompasses 8 languages, including Portuguese. This corpus can be used both for ASR and TTS (Text-To-Speech) model training. The Portuguese subcorpus contains 3.64 hours for training and 3.74 for test. The training set is composed from 26 male speakers and 16 female speakers voices, while the test set contains 5 male and 5 females. Common Voice Corpus⁸ [Wang et al. 2020] is a corpus created by the Mozilla Foundation. The release CoVoST 6.1 contains 63 hours of Portuguese (in the version pt_63h_2020-12-11), from which 50 hours were validated. The sexes are partially identified in the corpus, at least 81% of the speakers are males and at least 3% are females. The audios were donated by 1,120 speakers and uses the CC-0 license. The corpus was used to train the open model Deep Speech. By the time of writing the final version of this paper, the last version of CoVoST is version 11, with 150 hours, 125 validated. The Multilingual TEDx Corpus [Salesky et al. 2021] is composed of a collection of audio recordings from TEDx talks in 8 source languages, including 164 hours of Portuguese. This version is referred here as MTeDx-Pt.

CORAA ASR version 1.1 [Candido Junior et al. 2021] is a corpus with 290 hours of validated pairs (audio-transcription) extracted of five other corpora and adapted for the ASR task: (1) ALIP [Gonçalves 2019], composed of 78 hours from the interior areas of Brazilian state São Paulo; (2) C-ORAL Brasil I [Raso and Mello 2012], a corpus of 21 hours from Minas Gerais state; (3) Nurc-Recife [Oliveira Jr. 2016], which focuses on cultured urban norm and contains 279 hours from Recife city, Pernambuco; (4) SP2010 [Mendes and Oushiro 2012], presenting 65 hours of speech from the São Paulo city; (5) TeDx Talks in Portuguese, containing 72 hours from lectures and TeDx talks, different from those available in MTeDx-Pt. CORAA ASR was created in the scope of the Tarsila

⁶<https://igormq.github.io/datasets/>

⁷<https://librivox.org/>

⁸<https://commonvoice.mozilla.org/>

Project⁹, with both prepared and spontaneous speech, in contrast with the other corpora presented here, composed only of prepared speech.

2.2.3. Trained Models for Portuguese

This section presents four ASR models evaluated for the task of transcribing NURC/SP audios: [Candido Junior et al. 2021], [Ferreira and Oliveira 2022], [Grosman 2022], and [Stefanel Gris et al. 2022]. These works are based on Wav2vec 2.0 and were trained for Portuguese speech recognition using the corpora presented in Section 2.2.2.

[Candido Junior et al. 2021] trained the version XLR5-53 of Wav2vec 2.0 over CORAA ASR, reaching a WER of 24.18% for the mixed category (spontaneous and prepared speech). XLR5-53 is pretrained over several languages. The obtained value of WER seems higher compared to results in other works presented in this section. However, it is important to note that CORAA ASR is more challenging than other corpora, due to the presence of spontaneous speech and digitized audio. Besides that, CORAA ASR contains audios from NURC-Recife, which have similarities with NURC/SP. This model was chosen because it is closer to spontaneous speech audios in NURC/SP. [Ferreira and Oliveira 2022] trained different Wav2vec 2.0 XLR5-53 over different datasets. The authors explored domain specific ASR systems, mainly for prepared and spontaneous speech. In this work, we investigated the model built over CORAA ASR, CETUC, Common Voice 7.0, MLS and MTeDx-Pt. This model was exposed to more audios (prepared speech) than the model from [Candido Junior et al. 2021]. [Grosman 2022] trained Wav2vec 2.0 XLR-R over CORAA ASR, Common Voice 8.0, MTeDx and MLS. While the other models presented here were based on XLR5-53, this author used XLR-R, a bigger version of Wav2vec, with 1 billion parameters. The author reached a WER of 8.7% in Common Voice 8.0 using a language model for post-processing. This model was selected as because it is bigger, in number of parameters, than the other models presented here. We also explored the role of language model in the experiments involving this model. Finally, the work of [Stefanel Gris et al. 2022] is based on Wav2vec 2.0 XLR5-53 reaching 12.4% of WER evaluated against several large corpora in Portuguese, namely CETUC, MLS, Common Voice 7.0 and also some small corpora. This work was selected to contrast with the work models, since it has no spontaneous speech in its training set.

3. Experiments and Data Description

In this work, we compare the performance of the four ASR models described in Section 2.2.3, in a sample of five MC inquiries, totaling 5 hours 57 minutes and 33 seconds. The sample is small but representative as it varies in textual type and audio quality: three inquiries are type D2 (360, 396, 255), one inquiry is type EF (156) and another is type DID (242), shown in bold in Table 1. Transcripts of inquiries from the NURC/SP corpus were annotated according to the rules for oral transcription described in [Prete 1999]. Among the various phenomena annotated, in this article we evaluated the impact of three linguistic phenomena on the quality of automatic transcription of the evaluated models,

⁹<https://sites.google.com/view/tarsila-c4ai>

using WER/CER, in order to understand discrepancies in performance. The phenomena analyzed are: (1) “Incomprehension of words or segments”: annotated with “()”; (2) “Hypothesis of what was heard”: annotated with “(hypothesis)”; and (3) “Overlapping speakers’ voices”: annotated with “[”, followed by the speech snippet that was spoken at the same time as the speech before the square bracket. As a high error rate in an automatically generated transcript makes it difficult its manual revision, we expect from an ASR model that it generates transcripts in the WER range between $[0 .. 0,3)$, as this interval brings few errors that are easy to correct. In Table 2, we observe that the transcription rules indicate phenomena, such as “Overlapping speakers’ voices” (“[”) and “Incomprehension of words or segments” (“()”), that can impair the automatic recognition process, making the values of WER higher than expected.

Table 2. Examples of instances for the interval $0,3 \leq WER < 0,8$, using the annotation of the NURC project, for the SP_D2_360 inquiry. WER values for the examples appear between parentheses.

Annotated Transcripts /number of words	[Candido Junior et al. 2021]	[Ferreira and Oliveira 2022]	[Grosman 2022]
e dão muito trabalho tem esses esses problemas de juvenTUde esses negócios [() (não está muito na idade né?)/18	mue dão muito trabalho tem esse s problemas de juventude esses negócio não não uriamante ten (0,5)	me dão muito trabalho tem eos problemas de juventude esses negócio n não furiamunto tem (0,55)	que dá muito trabalho tem esses problemas de juventude esses negócios não uriamusic (0,44)
é eu [soube que também provocou um certo ciúmes ahn ahn isso eu (não) soube né eu VI... lá eu senti... um certo ciúmes ter:: éh ter sido escolhido uma mulher/31	e eu nsando que taambém provocou com certos filmes eu f isso eu soundão eu vi lá eu senti um certos filmes te tercido escolher de uma mulher (0,58)	é eu souo que teambém posocou com certos filmes isso eu soundão eu vi lá eu senti um certos filmes te ter sido escolher de uma mulher (0,54)	e eu me que terão e provocou um certo filmes e isso eu sou eu vi lá eu senti um certo ciúmes e ter sido escolhido uma mulher (0,38)

The script described in Section 3.1 is used to utterance units normalization of all inquiries analyzed in Section 3.2. This script which includes steps for dealing with the three linguistic phenomena presented above. The error analysis shown in Section 3.3 to assess inquiries’ performance using WER/CER is comparative and uses four sets of utterance units (in an ablation study) to assess the impact of the amount of voice overlap and the audio quality. Overlap is a characteristic phenomenon of spontaneous speech, which may help to explain low performances of ASR models. The audio quality, described with information about the **audio volume** (normal – audible, low, barely audible, loud) and **the quality of the recording** in relation to both the voice of the speakers (clear, good, medium bass, shaky voice) and to external events such as noise, hiss, and background music, is another factor which may influence the performance of the models.

3.1. Normalization Rules applied to MC Inquiries

As the annotations were kept to help in the segmentation of the utterances, a preprocessing was necessary to allow the speech segments to be processed by the ASR models evaluated in this article. The preprocessing of NURC inquiries is described below.

1. **Remove annotations:** this is done by removing only the snippets of utterance units that contain some snippet between double parentheses. For example, “hoje fui ((risos)) ao mercado” → “hoje fui ao mercado”;
2. **Remove extra characters:** removes all extra punctuation characters, parentheses (used to annotate “Incomprehension of words or segments” or “Hypothesis of what was heard”) and square brackets (used to annotate “Overlapping speakers’ voices”);
3. **Remove extra spaces,** e.g. “hoje fui ao mercado” → “hoje fui ao mercado”;
4. **Normalization:** normalizes filled pauses, removes hyphens, characters that do not belong to the alphabet, expands acronyms and cardinal and ordinal numbers, using the num2words library;
5. **Ignore empty utterance units:** this is done at the end because the normalization script can return empty utterance units.

It is important to note that: **ignore** here means not including the utterance unit in the test set, and **remove** means just treat the utterance unit (remove the problematic part).

3.2. Results using CER and WER Metrics

In Table 3 we show the performance of four ASRs trained with Brazilian Portuguese datasets, varying the amount of spontaneous and prepared speech data.

Models	SP_D2_255		SP_D2_360		SP_D2_396		SP_DID_242		SP_EF_156		All 21 inq.	
	WER	CER	WER	CER	WER	CER	WER	CER	WER	CER	WER	CER
[Candido Junior et al. 2021]	0,221	0,132	0,534	0,436	0,742	0,503	0,241	0,164	0,228	0,104	0,457	0,281
[Ferreira and dos Reis Oliveira 2022]	0,247	0,147	0,529	0,453	0,738	0,521	0,269	0,186	0,230	0,102	0,461	0,289
[Grosman 2022]	0,265	0,168	0,523	0,484	0,754	0,564	0,265	0,180	0,201	0,105	0,451	0,304
[Stefanel Gris et al. 2022]	0,333	0,186	0,656	0,516	0,859	0,600	0,319	0,203	0,321	0,150	0,588	0,365
<i>Average</i>	<i>0,267</i>	<i>0,158</i>	<i>0,561</i>	<i>0,472</i>	<i>0,773</i>	<i>0,547</i>	<i>0,274</i>	<i>0,183</i>	<i>0,245</i>	<i>0,115</i>	<i>0,489</i>	<i>0,309</i>

Table 3. Performance, using WER and CER, of four ASRs trained with BP data. We use the following order of colors in each column to rank the values of WER/CER: dark green (best), light green, yellow, orange, red (worst).

The ASR model described in [Candido Junior et al. 2021] presents the best average values of WER and CER for the five inquiries evaluated (0.393 and 0.268, respectively), while the model described in [Stefanel Gris et al. 2022], which was trained only with prepared speech, has the worst performance for all the five inquiries. [Ferreira and Oliveira 2022] and [Grosman 2022] models also presented consistent results, with a small advantage for Ferreira and Oliveira’s model. Considering that Grosman’s model is the only coupled with a language model, these results indicate that bigger models may require more data to fit for spontaneous speech.

Considering the average results of all 21 inquiries (last column in Table 3) the best values for WER and CER are Grosman’s model (0.451) and Candido’s model (0.281), respectively. We have chosen Candido’s to automatically transcribe the Audio Corpus of NURC/SP for three reasons. First, this model is faster. Second, CER is more reliable for

short audios analysis. Third, Candido’s CER is 2.3% better than Grossman’s CER, while Grossman’s WER is better, but by only 0.6% difference.

This evaluation has shown the need to include spontaneous speech in the training of the model since NURC/SP is composed of a large number of inquiries presenting informal dialogues between the speakers (D2) and interviews (DID). Furthermore, the spontaneous speech must be a considerable fraction of the corpus, otherwise the model can specialize in prepared speech. In order to better understand what have caused the increase of WER/CER values on the inquiries D2_360 and D2_396, we analysed two factors: the recording quality and the statistics of the linguistic phenomena annotated in the transcripts, because D2_360, for example, has characteristics similar to 30% of inquiries in the Audio Corpus, i.e. low or very low audio. We evaluated the impact of the quality of the audio files on the performance of the models, analysing one audio with good recording quality (clean sound) — D2_255, other with a low sound (D2_360) and another from the mixed class (D2_396), for two reasons: (1) we hypothesized that the voice overlapping would be largely responsible for the impact on performance, and (2) low audios could be correlated with the high rates of annotations of the type “Incomprehension of words or segments”, and “Hypothesis of what was heard”, thus impacting on the performance of the models. Table 4 shows the statistics of the annotations in the three inquiries of type D2.

Table 4. The correlation between the “Incomprehension of words or segments” and “Hypothesis of what was heard” with the audio quality: the values are low for the inquiry with good quality, but are high for the ones with mixed and bad quality.

Inquiries	Audio Quality	Overlapping	Incomprehension	Hypothesis
D2_360	-	1,215	89	348
D2_255	+	37	4	70
D2_396	Mixed	1,560	234	401

3.3. Error Analysis

In order to assess the impact of the amount of voice overlap and the audio quality on the performance of the ASR models evaluated, we created four sets of utterance units (called here case) for each inquiry. To create them, we processed the utterance units of the 21 inquiries of CM, with the best ASR evaluated in Section 3.2. It is expected that the Case 1, which has all the three linguistic phenomena of interest for this study, will have worst values of WER/CER (high values) and that the Case 4, which had the phenomena removed, will have the best values of WER/CER, i.e. low values:

- Case 1: contains utterance units with the three phenomena of interest: (i) “Incomprehension of words or segments”; (ii) “Hypothesis of what was heard”; (iii) “Overlapping speakers’ voices”.
- Case 2: utterance units with phenomena such as (i) and (ii) were excluded;
- Case 3: utterance units of type (iii) were excluded; and
- Case 4: utterance units of type (i), (ii) and (iii) were excluded.

In Figures 2(a) and 2(b) we show the variation of WER and CER, respectively, for each case presented above. The figure presents an ablation study to investigate the impact of the amount of voice overlap and the audio quality (shown in cases) on the WER/CER metrics.

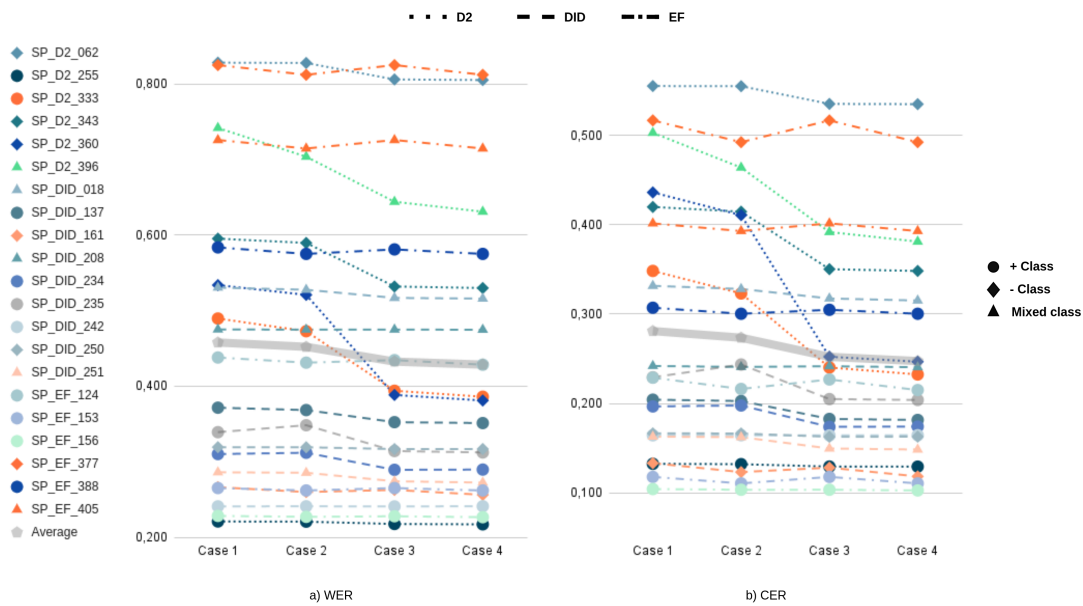


Figure 2. Figure 2(a) shows 11 inquiries showing the behaviour expected for the ablation experiment: there is a drop in WER values from Case 1 to Case 4, and for six of these inquiries of type D2 and DID the drop is accentuated (D2_396, D2_343, D2_360, D2_333, DID_235 and DID_234) as these type of inquiry presents informal dialogues between the speakers (D2) and interviews (DID). In contrast, for 10 inquiries, removing problematic utterance units does not affect ASR performance. For these, the performance of the ASR model is already good, with six inquiries showing WER values between 0.2 to 0.4 and four in the range of 0.4 to 0.6. Furthermore, four inquiries are of EF type, meaning a formal speech used in lectures and conference talks, where voice overlapping is rare. The two inquiries analysed in Section 3.2 (D2_396 and D2_360) presented an accentuated drop in WER and CER values (see Figure 2(b)) from Case 1 to Case 4.

4. Conclusions and Future Work

This work presented the NURC/SP corpus, discussing challenges to its preprocessing using ASR tools aiming at its public release. Particularly, we focused on the automatic transcription of 328 inquiries by evaluating four public available ASR systems in a manually aligned sample of NURC/SP. Our evaluation was performed using WER and CER metrics, as they are popular literature options for assessing ASR systems. Results suggest that [Candido Junior et al. 2021] model is more suitable for the proposed task. The model performance may be explained by it being fine-tuned for CORAA ASR instead of the proportion of spontaneous speech in its training set. Particularly, the NURC-Recife portion in CORAA ASR has similarities with NURC/SP that may be useful for models to identify spontaneous speech. The other ASR models evaluated here were also trained with other corpora having exclusively prepared speech, which can impact negatively on the performance of these models on spontaneous speech. In this work we also showed limitations of NURC/SP itself, namely: (1) “overlapping speakers’ voices” present in inquiries of types D2 and DID; (2) low audio quality in some of the inquiries, which impact even the manual transcription, with several annotations of “incomprehension of words or segments” and “hypothesis of what was heard”. As future work, we plan to fine-tune [Candido Junior et al. 2021] model using the 21 MC inquiries. In an iterative

approach, we intend to use this fine-tuned version to transcribe new inquiries and use the revised inquiries to feed back the fine-tuning process. We also plan to use a method for the manual revision of all 328 inquiries where the ASR scores for the presence of phonemes/graphemes in each audio segmentation will be used to determine the model confidence of each prediction. This way, segments with lower transcription confidence should be marked and reviewed more carefully by the annotators. This annotation will also be useful for linguistic studies and to allow finer grained audio selection of the corpus for ASR training.

Acknowledgements

This research was carried out at the Center for Artificial Intelligence (C4AI-USP), with support by the São Paulo Research Foundation (FAPESP grant #2019/07665-4) and by the IBM Corporation. V.G.S and F.R.F.S also thank, respectively, FAPESP for technical training fellowship grant #2020/16661-0 and CNPq for the research productivity grant #304961/2021-3.

References

- Alencar, V. F. S. and Alcaim, A. (2008). LSF and LPC - derived features for large vocabulary distributed continuous speech recognition in Brazilian Portuguese. In *2008 42nd Asilomar Conference on Signals, Systems and Computers*, pages 1237–1241.
- Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M. F., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates Inc.
- Candido Junior, A., Casanova, E., Soares, A., de Oliveira, F. S., Oliveira, L., Junior, R. C. F., da Silva, D. P. P., Fayet, F. G., Carlotto, B. B., Gris, L. R. S., and Aluísio, S. M. (2021). Coraa: a large corpus of spontaneous and prepared speech manually validated for speech recognition in brazilian portuguese. *arXiv preprint arXiv:2110.15731*.
- Ferreira, A. I. S. and Oliveira, G. d. R. (2022). Domain specific wav2vec 2.0 fine-tuning for the se&r 2022 challenge. In Marcacini, R., Junior, A. C., and Casanova, E., editors, *Proceedings of SE&R 2022, co-located with PROPOR 2022*, pages 9 – 14.
- Gonçalves, S. C. L. (2019). Projeto ALIP (Amostra Linguística do Interior Paulista) e banco de dados Iboruna: 10 anos de contribuição com a descrição do português brasileiro. *Revista Estudos Linguísticos*, 48(1):276–297.
- Grosman, J. (2022). XLS-R Wav2vec2 Portuguese by Jonas Grosman. <https://huggingface.co/jonatasgrosman/wav2vec2-xls-r-1b-portuguese>.
- Karpagavalli, S. and Chandra, E. (2016). A review on automatic speech recognition architecture and approaches. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 9(4):393–404.
- Li, J., Deng, L., Häb-Umbach, R., and Gong, Y. (2015). *Robust Automatic Speech Recognition: A Bridge to Practical Applications*. Elsevier Science.

- Mendes, R. B. and Oushiro, L. (2012). Mapping Paulistano Portuguese: the SP2010 Project. In *Proceedings of the VIIth GSCP International Conference: Speech and Corpora*, pages 459–463, Firenze, Italy. Firenze University Press.
- Oliveira Jr., M. (2016). NURC Digital: Um protocolo para a digitalização, anotação, arquivamento e disseminação do material do Projeto da Norma Urbana Linguística Culta (NURC). *CHIMERA: Revista de Corpus de Linguas Romances y Estudios Lingüísticos*, 3(2):149–174.
- Oliveira Jr., M. (2019). *NURC - 50 anos: 1969–2019*. Parábola Editorial, São Paulo, SP.
- Pratap, V., Xu, Q., Sriram, A., Synnaeve, G., and Collobert, R. (2020). MLS: A large-scale multilingual dataset for speech research. In Meng, H., Xu, B., and Zheng, T. F., editors, *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 2757–2761. ISCA.
- Preti, D. (1999). Normas para transcrição dos exemplos. In Preti, D., editor, *Análise de Textos Orais*, volume 1 of *Série Projetos Paralelos*, pages 11–12. Humanitas Publicações - FFLCH/USP, 4th edition.
- Raso, T. and Mello, H. (2012). *C-ORAL-BRASIL I: Corpus de referência do português brasileiro falado informal*. Editora UFMG, Belo Horizonte, MG.
- Salesky, E., Wiesner, M., Bremerman, J., Cattoni, R., Negri, M., Turchi, M., Oard, D. W., and Post, M. (2021). The multilingual tedx corpus for speech recognition and translation. *CoRR*, abs/2102.01757.
- Santos, V. G., Alves, C., Carlotto, B., Dias, B., Gris, L., Izaias, R., Morais, M. L., Oliveira, P., Sicoli, R., Svartman, F. R. F., Leite, M. Q., and Aluísio, S. (2022). CORAA NURC-SP Minimal Corpus: a manually annotated corpus of Brazilian Portuguese spontaneous speech. In *6th International Conference on Speech and Language Technologies on Iberian languages, IberSPEECH 2022, Granada, (to appear)*.
- Schneider, S., Baeovski, A., Collobert, R., and Auli, M. (2019). wav2vec: Unsupervised pre-training for speech recognition. In Kubin, G. and Kacic, Z., editors, *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, pages 3465–3469. ISCA.
- Stefanel Gris, L. R., Casanova, E., de Oliveira, F. S., da Silva Soares, A., and Candido Junior, A. (2022). Brazilian portuguese speech recognition using wav2vec 2.0. In Píneiro, V., Gamallo, P., Amaro, R., Scarton, C., Batista, F., Silva, D., Magro, C., and Pinto, H., editors, *Computational Processing of the Portuguese Language*, pages 333–343, Cham. Springer International Publishing.
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*. Curran Associates Inc.
- Wang, C., Pino, J., Wu, A., and Gu, J. (2020). CoVoST: A diverse multilingual speech-to-text translation corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4197–4203, Marseille, France. ELRA.