Automatic collection of transcribed speech for low resources languages

AGUIAR, Thales <http://orcid.org/0000-0002-1043-8685> and DA COSTA ABREU, Marjory <http://orcid.org/0000-0001-7461-7570>

Automatic Collection of Transcribed Speech for Low Resources Languages

1st Thales Aguiar
Department of Applied Math and Informathics
Federal University of Rio Grande do Norte
Natal, BR
0000-0002-1043-8685

2nd Márjory Da Costa-Abreu
Department of Computing
Sheffield Hallam University
Sheffield, UK
md0948@exchange.shu.ac.uk

***Abstract*—Speech is a crucial for human communication and combined with the evolution of instant messaging in voice format as well as automated chatbots, its importance is greater. While the majority of speech technologies have achieved high accuracy, they fail when tested for accents that deviate from the "standard" of a language. This becomes more concerning for languages that lack on datasets and have scarce literature, like Brazilian Portuguese. Thus, this paper proposes a methodology to collect and release a speech dataset for Brazilian Portuguese. The method explores the availability of data and information in video platforms, and automatically extracts the audio from TEDx Talks.**

***Index Terms*—dataset, TTBAcc, data collection, low resource, Brazilian Portuguese**

I. Introduction

With the increasing capacity and resources of computers and the rapid development of Artificial Intelligence (AI) methods, it has become possible to use them to impersonate other individual by Voice Conversion. However, those systems still are not effective on languages other than English, and even worse when it is a low-resource language [1]. Brazilian Portuguese (PT-BR) is an example, it suffers from a scarcity of speech data and research. For instance, to the best of our knowledge, there is no research about Accent Conversion (AC) for PT-BR. This is a more recent research topic, which aims to convert a source speech accent to a target speech accent, while preserving the identity of the source speaker.

Speech technologies are less concerned about the accents of a language. Instead, a single accent is used as a "one size fits them all", which is, generally, the most popular of a country. This excluding behaviour is harmful in an inclusive society, even more on a country with a continental size, like Brazil. For PT-BR, the scarcity of data regarding each accent, and in general, makes it difficult to use several of the current AI methods. Moreover, this situation also creates a challenge to creating speech systems and to speech research in PT-BR.

An accent is "*Modo de articular os sons, palavras ou frases, peculiar de uma pessoa, de uma região etc.*" (i.e., way to pronounce the sounds, words or sentences, specific to a person, region etc.) [2]. It represents a variation on how a population pronounces the phoneme across a language. Thus, it directly impacts on their communication and social interactions, therefore representing several cultural and social aspects of a person or population and making it a crucial part of what defines them [3]. The PT-BR carries historical facts, as the country has the presence of several cultures: Portuguese, African, Natives, Italian, and many others. Currently, there are at least 16 accents categorised by a group of Brazilian linguistics, the *Atlas Linguístico do Brasil*. Hence, preserving and understanding the accents, and the language itself, means to preserve the cultural aspects and the history of an ethnicity. The regions of the country that correponds to each accent is listed and illustrated in Fig. 1, while Fig. 2 illustrates the political regions of Brazil.

Fig. 1. Dialetical zones of Brazil.

A. Literature Review

This section briefly describes the most popular datasets for accent conversion. Librispeech [4] is often used for training intermediary models, such as acoustic models [5]–[7] or from pre-trained models [8], [9]. It has more than 1,000 hours of

Fig. 2. Political zones of Brazil.

English speech from 1,166 speakers, with pre-built language models and data for training language models.

L2-ARCTIC [10] is used by many works [5], [6], [11]–[13]. It is an English database with foreign speakers from 5 different languages using the CMU-ARCTIC [14] as stimuli. Each speaker has roughly 1,132 utterances (67±9 min), adding up to 11,032 utterances (11.2 h) in the dataset. The Voice Cloning Toolkit (VCTK) [15] was adopted for research in [12], [16]. It contains 46 hours from 106 speakers, and was collected to support the speaker adaptation popularity at the time.

Indic TIMIT [17] is an Indian to English speech dataset based on TIMIT prompts. It has 80 Indian speakers, from distinct regions, with 2,342 phonetically rich sentences in English. The reader can refer to Tab. II for a summary of the datasets.

There are two datasets available for PT-BR. The Common Voice [18] is a particularly big dataset for Portuguese, with 2,497 speakers and more than 116 hours of speech. However, it has no accent (a 'region' field is present, but almost no data is present) information and no indication of the Portuguese dialect. Another dataset is LapsBenchmark16k [19] with 700 sentences, 35 speakers, and approximately 54 minutes of audio. Although completely transcribed, the dataset is relatively small, and has no accent information.

Therefore, this research provides a new, larger, and public PT-BR speech dataset demographic information. This new dataset, TEDx Talks Brazilian Accents (TTBAcc), can be downloaded using the code available online [20].

II. Material and Methods

Inspired by the VoxCeleb [21], [22], this dataset also leverages the audio availability in video platforms. A source of mostly clean speech with several speakers and from multiple Brazilian states can be found in TEDx Talks. These are short videos, usually 8 or 15 minutes, with a defined subject and organised by independent communities. The speaker is often a public personality, and they make a brief presentation. Being a public personality, means that most of its personal information is public as well. This allows to collect such information from his social media, news, and other online resources.

The data is collected from a playlist [23] available on their channel, which has only Portuguese speakers. Nevertheless, other Portuguese dialects are easily identified or are found in the video description.

The collection process starts by automatically downloading the videos with its automatic and human-generated captions, whenever any are available, using a modified version of PyTube [24], [25]. This downloaded 4,276 videos, from which 324 did not have captions. There are 3,980 speakers discussing a variety of topics. To date, 800 samples were checked and 532 (520 valid) of those are PT-BR talks, summing up to 110 hours of annotated speech, from which 10.2 hours are annotated by humans. The audios were sampled at 44,1 kHz and have a resolution of 16 bit in MPEG format. Further, the audio is converted to WAV and resampled to 16 kHz.

Algorithm 1 Download and annotate data.

Require: $vidUrls.length > 0$
Ensure: Output result
$A \leftarrow [\,]$, $C \leftarrow [\,]$, $P \leftarrow [\,]$
$codec \leftarrow$ "pcm_s16le", $fr \leftarrow 16000$
$adModel \leftarrow$ "portuguese-mfa"
for $i \leftarrow 1$ to $vidUrls.length$ **do**
  $streams \leftarrow vidUrls[i].audioStreams()$
  if $len(streams) > 0$ **then**
    $cap \leftarrow$ downloadPtBRCaptions($streams$)
    $C.append(cap)$
    $stream \leftarrow$ last($streams$)
    $webm \leftarrow$ extractAudio($stream$, "webm")
    $wav \leftarrow$ FFMPEG2Wav($webm$, $codec$, fr)
    $D.append(wav)$
  end if
end for
$mfa \leftarrow$ trainMFA(ac: adModel, dict: adModel)
$textGrid \leftarrow mfa.align(D, C, adModel)$
$P.append(textGrid)$
return (A, C, P)

The Algorithm 1 returns a dataset $D = (A, C, P)$, for a collection of WAV audios A, a collection of captions C, and a collection of phonetic annotations P.

Demographics are unimportant for AC, but are crucial for other speech tasks and can push forward the PT-BR research. To collect this information, we first access the video URL and check if there is more than one speaker, speech disorders, singing, or musical instruments. If any, the sample is invalid. Otherwise, the gender of the speaker is annotated and a manual search begins. Any online resource about the speaker is used, news, social media, and any curriculum. Then, the *age*,

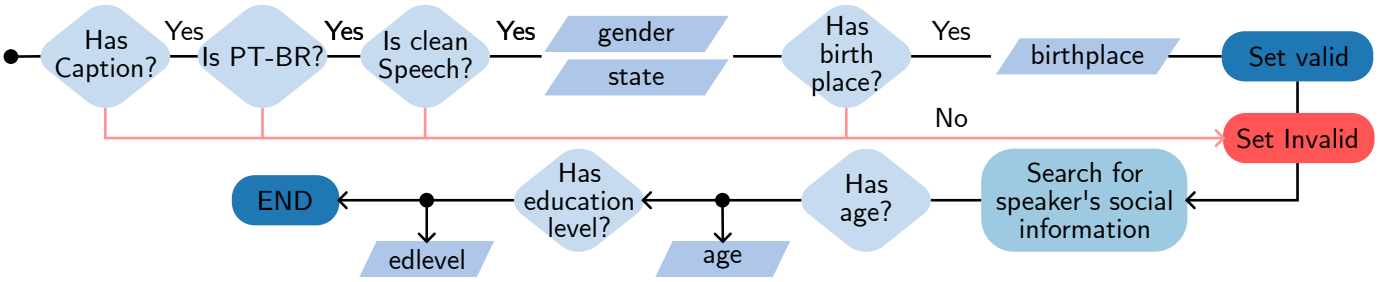

Fig. 3. Dataset validation steps. "Clean Speech" means there are no speech disorders, singing, music, musical instruments, or multiple speakers. Red path means the answer to any of the questions is "No".

education level, *state* and *birthplace* are extracted from any of the sources. Age can be inferred by information crossing. For instance, one source has the passage "graduate of college at 21", while other tells him graduated from college in 2018.

Since the aim of the dataset is enabling AC for PT-BR, the crucial features of the dataset are the **state** and **birthplace** of the speaker. When the search does not provide the state where the speaker was born, this sample is marked as invalid since without this, it is impossible to know which accent is being converted. After finding the birthplace, it is important to check if the speaker still lives at this location. If positive, then the birthplace feature is set to `True`, otherwise to `False`. The complete process is illustrated in Fig. 3. Although invalid, a sample may still be useful for speech tasks besides AC.

After collecting and validating, the dataset is forced-aligned, generating the time intervals for each word and phoneme in their captions. This process is performed using the Montreal Forced Aligner. Then, the dataset quality is measured regarding the audio and captions. The first is measured by the voiced duration of the talks, and the spectral flatness. The automatic generated captions have their quality tested using samples that have both automatic and human captions. By considering the human caption as the gold-standard and the automatic caption as the prediction, it is possible to compute the Word Error Rate (WER), Mean Error Rate (MER) and Word Information Loss (WIL) and get a notion of the automatic caption quality.

III. TTBACC DATASET

This section describes and provides a visualisation of the TTBACC dataset. All information given here is considering valid PT-BR samples (**18.7% of all samples**) to date.

A. Dataset description

The dataset is well distributed between male and female speakers (Tab. I). For states (Fig. 2), on the other hand, one gender can become highly dominant, like *Rio Grande do Norte*. However, this behaviour is observed in regions that have a few speakers (Fig. I). Thus, this inequality is caused by a regional under-representation from the TTBAcc validation state.

The most difficult feature to collect was the age, unavailable in 55.19% of the samples. Frequently, age was available when the speaker is a more popular figure such as writers, actors, businessmen, and politicians. Then, the birthplace was the second most difficult to find. Although somewhat dispensable, it may help when screening for accents, as a more accurate location of the speaker facilitates to associate him with the correct accent. Following, the education level was unavailable for 10.19% of the speakers. However, those cases are commonly from speakers that do not have much information online. Finally, there were no captions for 0.38% (3 out of 800) of the talks. This happened in the automatic downloading process, and is controlled by the owners of the videos. For the 518 captions, 70 (13.51%) had captions annotated by humans and 448 (86.49%) have only automatically generated captions.

Currently, the TTBAcc covers 21 out of the 27 (77.78%) Brazilian states. The exceptions are *Roraima, Amapá, Rondônia, Tocantins, Piauí*, and *Maranhão*. The data is concentrated in the wealthier areas of the country: the South and South-east. Although, every region is represented, some has much fewer data, e.g., North has 4 speakers in total (Tab. I). However, considering there are over 3,000 samples to be validated, the level of representation is expected to increase.

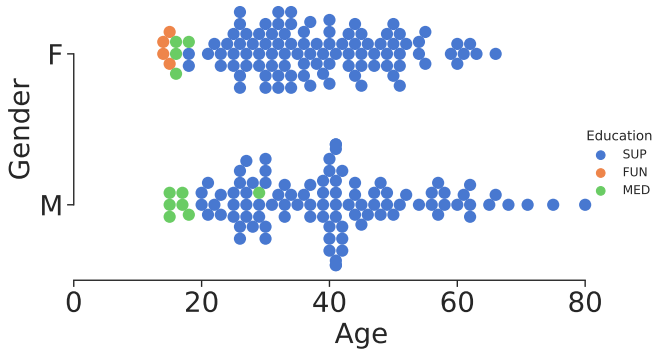

Fig. 4. Distribution of gender and educational level per state.

The age of the speakers between states has a tight average, considering data in Tab. I while the overall mean age of the dataset suggests a predominance of younger speakers. Fig. 4 illustrates the age distribution regarding other dataset variables. The first compares the age to the educational level and the gender of the speaker. While only women are representing the FUN education level, only a few speakers are over 20

TABLE I

TTBACC DESCRIPTION PER DIALECT. ONLY THE PREVALENT ACCENTS ARE LISTED. AGE IS DESCRIBED AS MEAN ± STANDARD DEVIATION. AT THE BOTTOM, TOTALS PRESENT THE SPEAKER COUNT, OVERALL AGE MEAN AND STANDARD DEVIATION, AND THE DATASET SPEECH DURATION IN MINUTES.

Accent	State	Gender		Education			Age (years)	Length (min)
		M	F	S	M	F		
Nortista	*Acre* (AC)	1	0	1	0	0	—	3.63
	Amazonas (AM)	0	1	1	0	0	—	19.30
	Pará (PA)	0	1	1	0	0	—	10.88
Nordestino, *Baiano*, *Costa Norte*, *Recifense*	*Alagoas* (AL)	2	0	1	0	0	—	20.15
	Bahia (BA)	11	18	26	0	0	34.25±8.84	345.25
	Ceará (CE)	0	1	1	0	0	—	9.37
	Distrito Federal (DF)	3	3	5	0	0	35.20±9.54	94.42
	Paraíba (PB)	2	0	2	0	0	—	30.83
	Pernambuco (PE)	13	3	13	0	1	39.83±9.91	209.13
	Rio Grande do Norte (RN)	0	1	1	0	0	—	5.50
	Sergipe (SE)	0	1	1	0	0	—	18.17
Sertanejo, *S.Amazônica*, *Brasiliense*	*Goiás* (GO)	1	0	1	0	0	—	16.07
	Mato Grosso do Sul (MS)	0	2	2	0	0	—	20.01
	Mato Grosso (MT)	4	2	5	0	0	37.50±10.61	86.08
Mineiro, *Paulistano*, *Caipira*, *Fluminense*	*Minas Gerais* (MG)	32	39	63	0	0	36.22±12.58	954.98
	Espírito Santo (ES)	1	3	4	0	0	—	64.78
	Rio de Janeiro (RJ)	35	33	59	2	0	38.43±12.48	902.50
	São Paulo (SP)	89	102	162	9	4	36.32±14.33	2181.93
Sulista, *Gaúcho*	*Paraná* (PR)	28	22	46	0	0	45.15±12.46	704.43
	Rio Grande do Sul (RS)	15	24	33	1	0	42.39±12.74	526.93
	Santa Catarina (SC)	14	13	20	2	0	48.28±23.37	363.42
Total	—	251	269	448	14	5	38.17±13.69	6587.34

years and have the MED education level. The remainder of speakers is graduated or post-graduated level. With the majority speakers close to 40 years for men, and between 30 and 40 for women. Therefore, showing that a good split for comparison between genders would be in this range. However, this graph shows the predominance regarding the number of samples. People with basic education (FUN and MED) have less speech data, but with a good length variation. Moreover, similar to Fig. 4, the speakers with basic instructions are few below the thirties.

TABLE II

POPULAR DATASETS FOR ACCENT CONVERSION. TEXT INSIDE **[]** IS THE TARGET LANGUAGE. 'SPKS' IS FOR UNIQUE SPEAKERS.

Name	Hours	Languages	Spks
L2-ARCTIC [10]	11.2	IN, KO, CN, SP, AR **[EN]**	10
Indic TIMIT [17]	240.0	IN **[EN]**	80
VCTK [15]	≥300.0	EN-UK **[EN-UK]**	500
Common Voice [18]	140.0	PT	2,967
LapsBM16k [26]	0.9	PT-BR	35
TTBAcc [20]	110.0	PT-BR	455

B. Phonetic Annotation

A phonetically labelled dataset is another resource that PT-BR lacks of. However, manually aligning and labelling phonemes is a laborious work that requires specialists analysing tiny audio segments, several times. Therefore, this paper uses the MFA [27] to automatically annotate the phonemes of the dataset. The tool align the phonemes to their respective intervals, allowing future works to perform other phoneme related tasks. For MFA to work, it requires a set of audio, their captions, a phonetic dictionary, and an acoustic model. Both the dictionary and acoustic model for PT-BR are available in their list of supported languages. There are two versions of dictionaries and models, this work uses the `portuguese-mfa` versions. The process aligned 458 talks, remaining talks presented problems with parameter tuning or out-of-vocabulary words.

IV. QUALITY OF DATA

Having a large dataset is important, but the quality of the data is also crucial. The quality of the collected data is evaluated regarding speech and the transcription.

A. Speech

The TTBAcc has a variety of environments where speakers make their presentations. Those places can have variations in sound quality and recording devices. Therefore, precisely estimating the silence duration is difficult. To have an approximation of such a metric, we applied the function from Librosa [28] to separate voiced and unvoiced segments of an audio. The algorithm estimates the background noise and takes a user-defined threshold in dB to identify silence sections. The audio from TTBAcc are segmented using 60, 45, and 30 dB. The results in Fig. 5 suggest there are from 3.20 h to 12.84 h

of silence in the dataset. Since 32.53 h is too much considering the author perception during validation, this result is discarded.

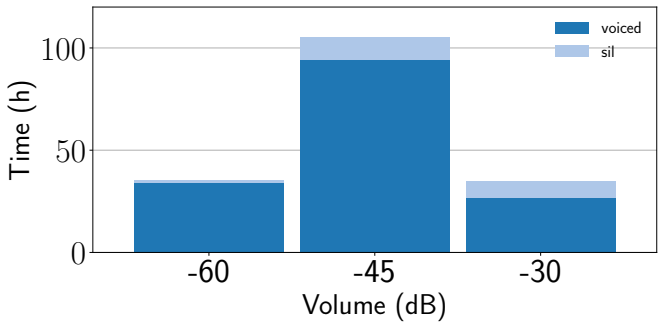

Fig. 5. Time, in hours, of voiced and silence duration of the dataset.

The dataset has approximately 109.8 h, and 12.84 h of silence in the worst scenario. Therefore, there is at least 96.94 h of voiced audio in TTBAcc. With an 11.7% estimated silence in the dataset. Thus, most of the audio is meaningful.

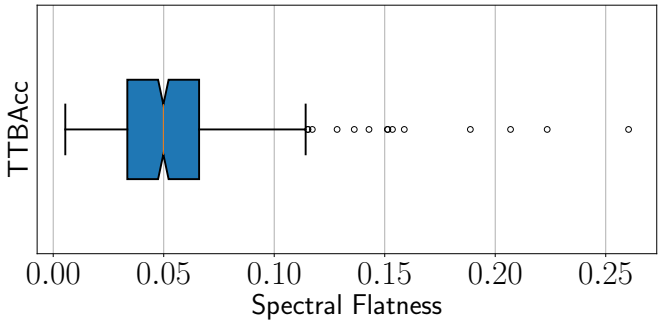

Fig. 6. Spectral flatness distribution of ttbacc.

Spectral Flatness (w) [29] is another quality metric. It computes a similarity value between the signal and white noise, with 0 meaning no similarity and 1 total similarity. Fig. 6 presents the distribution of the spectral flatness of TTBAcc. It was obtained with the same data used for silence estimation. It also shows that most of our signals (**Mdn.** w = 0.050), corresponds to a low similarity with white noise. Although this does not ensure a high quality of the speech signal, it is a good suggestion.

B. Transcription

Some videos have both human and machine transcriptions. Therefore, for measuring the quality of the collected dataset, computing the WER, MER, and WIL of those samples can provide an insight about its quality. Fig. 7 displays the WER, MER, and WIL computed with the JiWER package [30]. Both *gold-standard* and the AI-generated captions are transformed with the following operations: lower case, remove punctuation, white spaces, and Kaldi non-words. The last one removes words between `[]` or `()` like `[laugh]`.

The results are calculated from 70 samples (10.2 hours). WER and MER presents similar distributions, the former has

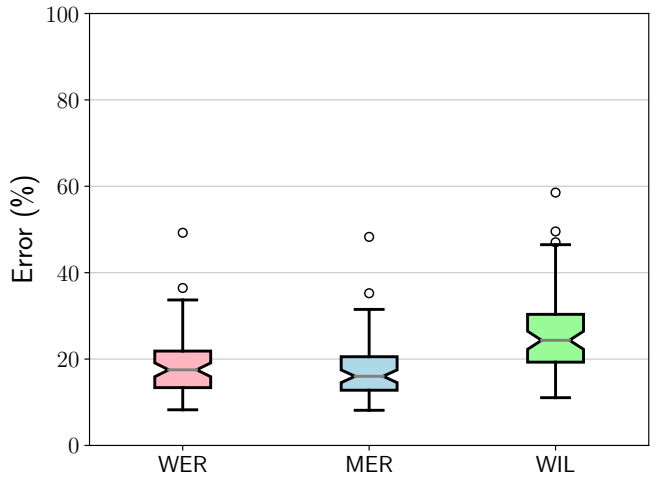

Fig. 7. TTBAcc caption quality in WER, MER, and WIL. Measures obtained from 10.2 hours of speech.

a mean error rate of 18.87% ± 7.53% and Mdn. of 17.51% while the latter has 17.90% ± 7.05% with a Mdn. of 16.01%. Furthermore, both have two outliers. Since MER only adds an upper limit to the WER, the close results are expected. However, WIL has a higher mean and median: 26.56% ± 11.06% and 24.53%, respectively. This metric will often be higher, as [31] shows on table 1 of their work. Moreover, the notches from WIL do not overlap with neither WER nor MER. Thus, there is a significant difference in their medians [32].

A human annotator achieves ≈ 4.5% WER, while the automatic captions are likely to have 14% more errors. Given this significant difference, AC usage may be restricted to human annotated captions. However, experiments must be performed to confirm the viability of using machine transcripts.

V. Conclusions and Limitations

The education level of a person can have a considerable impact on its accent [33]. Often, a higher education level means the accent of the individual will shift towards the "standard". Therefore, diversity of education level will capture more of an accent than a collection of highly educated individuals. Since the TEDx Talks are usually presented by scholars and businesspersons, most of our data contains speech from individuals with a superior education level, restricting the coverage of an accent.

The methodology presented in this work cannot account for precise accent. Therefore, the only information that one can obtain automatically is the region, state, and birthplace of a speaker. Although these data not necessarily indicates the accent of a person, there is no other way of collecting data for accent conversion without a specialist. A few problems may arise from this method, since one can change (or lose) its accent when living in another place for too long. Furthermore, the political boundaries do not reflect the linguistic boundaries of a population. Such lurking variable cannot be addressed in

our methodology, and is accepted as a trade-off between the precision and quantity of data.

REFERENCES

[1] T. Aguiar de Lima and M. Da Costa-Abreu, "A survey on automatic speech recognition systems for Portuguese language and its variations," *Comput. Speech Lang.*, vol. 62, p. 101055, Jul. 2020.

[2] Michaelis, "Definition of "sotaque"," Dicionário Brasileiro da Língua Portuguesa, Mar. 2022.

[3] D. Callou, *Iniciação à fonética e à fonologia*. Rio de Janeiro: Jorge Zahar, 1990.

[4] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: An ASR corpus based on public domain audio books," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2015, pp. 5206–5210.

[5] G. Zhao, S. Ding, and R. Gutierrez-Osuna, "Foreign accent conversion by synthesizing speech from phonetic posteriorgrams," 2019, pp. 2843–2847.

[6] A. Das, G. Zhao, J. Levis, E. Chukharev-Hudilainen, and R. Gutierrez-Osuna, "Understanding the effect of voice quality and accent on talker similarity," 2020, pp. 1763–1767.

[7] S. Liu, D. Wang, Y. Cao, L. Sun, X. Wu, S. Kang, Z. Wu, X. Liu, D. Su, D. Yu, and H. Meng, "End-To-End Accent Conversion Without Using Native Utterances," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing*, 2020, pp. 6289–6293.

[8] G. Zhao, S. Sonsaat, J. Levis, E. Chukharev-Hudilainen, and R. Gutierrez-Osuna, "Accent Conversion Using Phonetic Posteriorgrams," 2018, pp. 5314–5318.

[9] G. Zhao and R. Gutierrez-Osuna, "Using Phonetic Posteriorgram Based Frame Pairing for Segmental Accent Conversion," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 27, no. 10, pp. 1649–1660, 2019.

[10] G. Zhao, S. Sonsaat, A. Silpachai, I. Lucic, E. Chukharev-Hudilainen, J. Levis, and R. Gutierrez-Osuna, "L2-Arctic: A non-native English speech corpus," 2018, pp. 2783–2787.

[11] G. Zhao, S. Ding, and R. Gutierrez-Osuna, "Converting Foreign Accent Speech without a Reference," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 29, pp. 2367–2381, 2021.

[12] D. Wang, S. Liu, L. Sun, X. Wu, X. Liu, and H. Meng, "Learning explicit prosody models and deep speaker embeddings for atypical voice conversion," vol. 4, 2021, pp. 3031–3035.

[13] C. Liberatore, "Native-nonnative voice conversion by residual warping in a sparse, anchor-based representation," *IEEE/ACM Transactions on Audio Speech and Language Processing*, vol. 29, pp. 3040–3051, 2021.

[14] J. Kominek and A. Black, "The CMU Arctic speech databases," *SSW5-2004*, Jan. 2004.

[15] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *2013 International Conference Oriental COCOSDA held jointly with 2013 Conference on Asian Spoken Language Research and Evaluation (O-COCOSDA/CASLRE)*, 2013, pp. 1–4.

[16] D.-Y. Wu, Y.-H. Chen, and H.-Y. Lee, "VQVC+: One-shot voice conversion by vector quantization and U-Net architecture," 2020, pp. 4691–4695.

[17] C. Yarra, R. Aggarwal, A. Rajpal, and P. K. Ghosh, "Indic TIMIT and Indic English lexicon: A speech database of Indian speakers using TIMIT stimuli and a lexicon from their mispronunciations," 2019.

[18] Mozilla, "Common voice portuguese dataset," May 2022, accessed on the 04th May. [Online]. Available: https://commonvoice.mozilla.org/pt

[19] N. C. S. Neto, "Ferramentas e recursos livres para reconhecimento e síntese de voz em português brasileiro," phdthesis, Federal University of Pará, Instituto de Tecnologia, Jun. 2011.

[20] T. A. de Lima, "Tedx talks brazilian accent dataset," May 2023. [Online]. Available: https://github.com/thalesaguiar21/accent-dataset

[21] J. S. Chung, A. Nagrani, and A. Zisserman, "VoxCeleb2: Deep Speaker Recognition," in *Proc. of the Interspeech*. Hyderabad, India: ISCA, Sep. 2018, pp. 1086–1090.

[22] A. Nagrani, J. S. Chung, and A. Zisserman, "VoxCeleb: A Large-Scale Speaker Identification Dataset," in *Proc. Interspeech 2017*, 2017, pp. 2616–2620.

[23] TED, "Playlist of tedx talks in brazilian portuguese," May 2022, accesed on the 22nd of May. [Online]. Available: https://bit.ly/3PTZbak

[24] ehsanbehdad, "Fixed version of pytube to download videos," May 2022, accesed on the 09th of May. [Online]. Available: https://github.com/ehsanbehdad/pytube

[25] Y. Miklin, "Fixed version of pytube to download captions," May 2022, accesed on the 12th of May. [Online]. Available: https://stackoverflow.com/a/69004322

[26] FalaBrasil, "LapsBenchmark 16k repository," Sep. 2018, accessed 17 December 2019.

[27] M. McAuliffe, M. Socolof, S. Mihuc, M. Wagner, and M. Sonderegger, "Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi," in *Proc. of the Interspeech*. Stockholm, Sweden: ISCA, Aug. 2017, pp. 498–502.

[28] B. McFee, A. Metsai, M. McVicar, S. Balke, C. Thomé, C. Raffel, F. Zalkow, A. Malek, D. , K. Lee, O. Nieto, D. Ellis, J. Mason, E. Battenberg, S. Seyfarth, R. Yamamoto, Viktorandreevichmorozov, K. Choi, J. Moore, R. Bittner, S. Hidaka, Z. Wei, Nullmightybofo, A. Weiss, D. Hereñú, F.-R. Stöter, P. Friesch, M. Vollrath, T. Kim, and T. , "librosa/librosa: 0.9.1," 2022.

[29] J. Johnston, "Transform coding of audio signals using perceptual noise criteria," *IEEE Journal on Selected Areas in Communications*, vol. 6, no. 2, pp. 314–323, 1988.

[30] jitsi, "Jiwer: Similarity measures for automatic speech recognition evaluation," Jun. 2022, v 2.3.0, Accessed on the 9th of June. [Online]. Available: https://github.com/jitsi/jiwer

[31] R. Errattahi, A. El Hannani, and H. Ouahmane, "Automatic Speech Recognition Errors Detection and Correction: A Review," *Procedia Comput. Sci.*, vol. 128, pp. 32–37, 2018, 1st International Conference on Natural Language and Speech Processing.

[32] R. McGill, J. W. Tukey, and W. A. Larsen, "Variations of box plots," *The american statistician*, vol. 32, no. 1, pp. 12–16, 1978.

[33] B. Cowan, H. Branigan, P. Doyle, J. Edwards, D. Garaialde, J. Cabral, A. Hayes-Brady, and L. Clark, "What's in an accent? The impact of accented synthetic speech on lexical choice in human-machine dialogue." Association for Computing Machinery, 2019.