EUROPEAN LANGUAGE EQUALITY 2

FSTP Project Report

pangeanic-voice – Generation of a Large Speech Corpus for the Languages of Spain using Data Augmentation

Authors	José Miguel Herrera, Moisés Barrios
Organisation	Pangeanic
Dissemination level	Public
Date	30-03-2023

About this document

Project	European Language Equality 2 (ELE2)
Grant agreement no.	LC-01884166 – 101075356 ELE2
Coordinator	Prof. Dr. Andy Way (DCU)
Co-coordinator	Prof. Dr. Georg Rehm (DFKI)
Start date, duration	01-07-2022, 12 months
FSTP Project	pangeanic-voice – Generation of a Large Speech Corpus for the Languages of Spain using Data Augmentation
Authors	José Miguel Herrera, Moisés Barrios
Organisation	Pangeanic
Type	Report
Number of pages	22
Status and version	Final
Dissemination level	Public
Date of delivery	30-03-2023
EC project officer	Susan Fraser
Contact	European Language Equality 2 (ELE2) ADAPT Centre, Dublin City University Glasnevin, Dublin 9, Ireland Prof. Dr. Andy Way – andy.way@adaptcentre.ie European Language Equality 2 (ELE2) DFKI GmbH Alt-Moabit 91c, 10559 Berlin, Germany Prof. Dr. Georg Rehm – georg.rehm@dfki.de http://www.european-language-equality.eu © 2023 ELE2 Consortium

Consortium

1	Dublin City University (Coordinator)	DCU	IE
2	Deutsches Forschungszentrum für Künstliche Intelligenz GmbH (Co-coordinator)	DFKI	DE
3	Univerzita Karlova (Charles University)	CUNI	CZ
4	Universidad Del Pais Vasco/ Euskal Herriko Unibertsitatea (University of the Basque Country)	UPV/EHU	ES
5	Athina-Erevnitiko Kentro Kainotomias Stis Technologies Tis Pliroforias, Ton Epikoinonion Kai Tis Gnosis	ILSP	GR
6	European Federation of National Institutes for Language	EFNIL	LU
7	Réseau européen pour l’égalité des langues (European Language Equality Network)	ELEN	FR

Contents

List of Figures

List of Tables

List of Acronyms

AI	Artificial Intelligence
AI4EU	AI4EU (EU project, 2019-2021)
CLAIRE	Confederation of Laboratories for AI Research in Europe
CLARIN	Common Language Resources and Technology Infrastructure
CRACKER	Cracking the Language Barrier (EU project, 2015–2017)
DARIAH	Digital Research Infrastructure for the Arts and Humanities
DLE	Digital Language Equality
EC	European Commission
ECSPM	European Civil Society Platform for Multilingualism
EFNIL	European Federation of National Institutes for Language
ELE	European Language Equality
ELE2	European Language Equality *(this project)*
ELE Programme	European Language Equality Programme *(the long-term, large-scale funding programme specified by the ELE project)*
ELEN	European Language Equality Network
ELEXIS	European Lexicographic Infrastructure
ELG	European Language Grid (EU project, 2019-2022)
ELRA	European Language Resource Association
ELRC	European Language Resource Coordination
ELT	European Language Technology
EP	European Parliament
ERIC	European Research Infrastructure Consortium
ESCO	European Skills, Competences, Qualifications and Occupations classification

GDPR	General Data Protection Regulation
KPI	Key Performance Indicator
LT	Language Technology/Technologies
META	Multilingual Europe Technology Alliance
META-NET	EU Network of Excellence to foster META
ML	Machine Learning
MT	Machine Translation
NCC	National Competence Centre
NCP	National Contact Point
NLP	Natural Language Processing
STOA	Science and Technology Options Assessment

Abstract

As virtual assistants and smart home devices become increasingly common in our daily lives, speech recognition systems are also becoming more prevalent. Unfortunately, there is currently a lack of comprehensive datasets for EU languages in the field of speech technology. However, data augmentation techniques can play a vital role in creating an extensive and diverse dataset.

To contribute to the ELE strategic agenda, in this work we create a guideline for building an extensive speech dataset with transcriptions of languages spoken in Spain through audio data augmentation (ADA) techniques. By incorporating ADA techniques such as noisy backgrounds, time masking, and speed variation, we aim to expand a standard dataset into a much larger speech dataset, by a factor of at least 20. This approach provides a better representation of various speech and sound types, as it simulates real-life environments.

1 Introduction

Speech recognition technology enables us to transcribe spoken audio into text, eliminating the need for manual writing on a device. With this technology, users can dictate emails, respond to text messages, transcribe documents, and control smart home devices, among other things. This technology is useful in various fields. For instance, in the healthcare industry, it can assist individuals with physical difficulties in writing. It can also be used in tasks that require transcribing a recorded conversation to perform text analysis, such as speeches.

To successfully implement this technology, a substantial amount of speech data and the corresponding transcriptions are necessary for model training. However, a recent ELE report Consortium, 2022 reveals that English has the highest level of technological support, followed by German, Spanish, and French, with less than half the level of support compared to English.

There is limited or potentially no support for other languages or dialects spoken in European Union (EU) countries. Furthermore, existing datasets suffer from limited representation in terms of gender, age range, or foreign accents, and they do not account for real-world scenarios like background noise. Training speech recognition models requires numerous audio segments along with their corresponding transcriptions, which is a challenging task that requires many hours of recording individuals. Therefore, it is crucial to collect more diverse and comprehensive datasets to improve the accuracy and inclusiveness of the technology for a broader user base. In this work, we provide a guideline for building a speech dataset with transcriptions using data augmentation (DA) techniques. It means that for a given small dataset, we can add several types of real noises such as streets, stadiums, screaming children, and subways, among others. Moreover, we can vary the audio by altering the speed or adding some distortions in the speech. With these changes and considering diverse types of speech (gender and age) it is possible to increase the initial dataset.

This approach could be extended to any language spoken in the EU, but we focus our attention on Spanish languages. In particular, Euskera, Catalan, Galician, and Asturian languages. To achieve this, we enlist individuals to record several texts in their native language (compensating them accordingly). Figure 1 shows one instance of our approach.

This report outlines first an overview of ADA and the techniques that we used in section 4. Then, we apply a comprehensive methodology for implementing and extending a speech project, followed by a real case study that demonstrates the methodology's effectiveness in transforming a limited number of audio files into several variations.

Through this report, we hope to encourage the speech technology community to generate more and larger datasets in other EU languages. Moreover, this proposal is in direct line with the strategic agenda of ELE because it provides a large dataset with speech transcriptions and

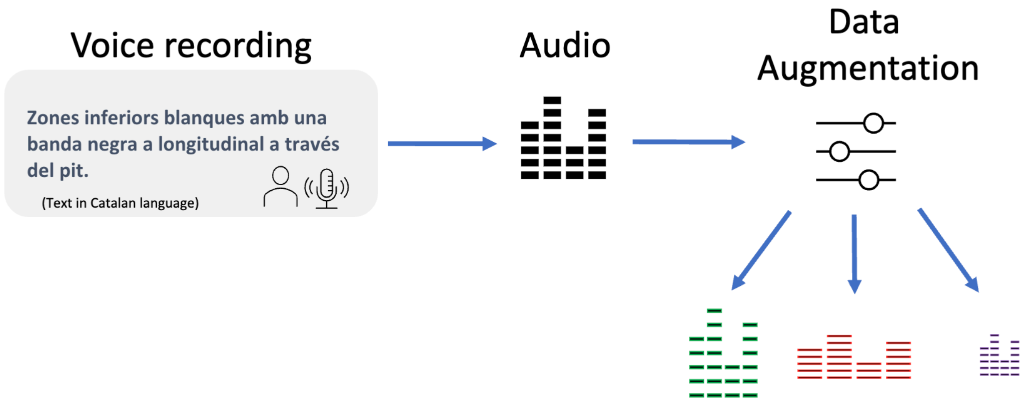

Voice recording
Zones inferiors blanques amb una banda negra a longitudinal a través del pit.
(Text in Catalan language)
Audio
Data Augmentation

Figure 1: Given a text in a specific language (in this case Catalan), we reach out to fluent speakers and ask them to record. This recording is then processed using ADA techniques to increase the size of our data set by at least 20 times.

simulates real audio environments.

This work has led us to share $63,271$ audio files of speech, each with its own transcription and metadata. This is equivalent to 200 hours of speech without ADA. Additionally, we've made the ADA source code available. While this dataset may not suffice for training a speech recognition model, the guidance we offer will prove invaluable in developing more comprehensive speech projects on a larger scale.

2 Audio Data Augmentation (ADA)

In this section, we provide an overview of ADA and the techniques we used to carry out this project.

ADA is a method used to increase the size of a dataset by creating new examples through modifications of existing data. This technique is commonly used in machine learning and computer vision tasks to enhance model performance by introducing variations in the input data. In our context, ADA involves applying a series of transformations to expand an existing dataset. This process entails altering audio by introducing background noise, distortion, speed changes, and other modifications. Figure 2 provides one example of ADA, where the original audio file is altered by adding an airport noise and a time stretch transformation (increase or decrease speech speed).

Generating a speech dataset is a demanding task as it requires human effort. Collecting a large dataset with varied speech samples can be both time-consuming and expensive. ML models rely on clean and high-quality audio files for various purposes, such as transcriptions. In addition, having insufficient data can be a major challenge in building and improving models. ADA techniques can be particularly useful for audio files as they can help in addressing the lack of data. By augmenting the existing dataset, we can generate a more extensive and diverse set of audio samples, which can improve the performance of the models. ADA can also help to reduce the economic and time costs of data collection, making it a more feasible approach. Furthermore, augmenting the data can introduce diversity in the dataset, which can lead to better model generalization and improved accuracy.

We experimented with different techniques (Ferreira-Paiva et al., 2021), such as pitch shifting, room simulation, time masking, and various filtering methods. Finally, we selected

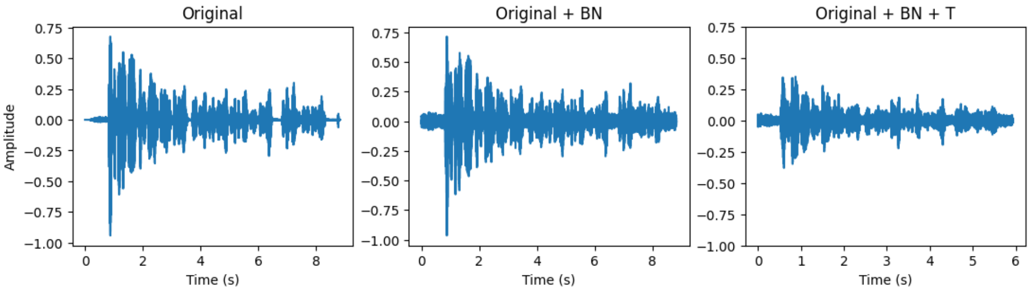


Figure 2: The effect of ADA when we apply an airport background noise (BN) and then speed up (T) the original audio file.

four techniques that we believe are representative. We implement them using **audiomentations**[1]. Although there are several libraries available for ADA, the **audiomentations** library stands out as a comprehensive tool with numerous techniques that are both user-friendly and well-documented. It is straightforward to use and implement, with clear documentation and examples available to guide users through the process.

The parameters that we used for each technique are explained briefly below.

- **Background Noise**: It adds background noise from a set of audio files with different types of noise (Salamon and Bello, 2017).

  We use the following parameters:

  1. The **Signal-to-Noise Ratio (SNR)** in decibels (dB), where lower values indicate that the noise is being amplified. Since it is measured in dB, it can take both positive and negative values.
  2. Whether the **Root Mean Square (RMS)** of the added noise should be proportional (relative) or independent (absolute) of the RMS of the input sound. In particular, we set the parameter to relative.

  To carry out this approach, it was necessary to obtain a set of background audio files. For this task, we extracted several hours of each type of real environment sound as airport noise, crying babies, city noise, construction noise, rain, children, stadium noise, subway noise, and traffic noise. Then, we generated new files from random segments of these audios, between 3 and 14 seconds long. Despite acquiring a large number of background audio segments, we decided to use 22 samples of each noise type to ensure an equal probability of selection for each. As a result, we utilized a total of 198 background noise files in our ADA process.

- **Gaussian Noise**: it simulates certain ambient sounds like the hum of an air conditioner or the buzz of a television. This method involves introducing white noise to the audio signal, which is generated from a normal distribution or Gaussian distribution (normal distribution).

  We use the following parameters:

  1. The **amplitude factor** indicates the amount of added noise that can simulate anything from a low-quality recording or low background noise to a recording in extreme noise conditions or an audio signal with very high ambient noise.

[1] https://github.com/iver56/audiomentations

- **Tanh Distortion**: It emulates the distortions of the speech. The effect imitates speech distortions by utilizing the hyperbolic tangent activation function on the input audio, causing speech to become distorted.

  We use the following parameters:

  1. The **distortion level** determines the amount of harmonic distortion applied to the signal. A higher value results in more significant distortion, while a lower value will produce less distortion.

- **Time Stretch**: It increases or decreases the speed of an audio signal without changing its pitch or intonation. The timing of the audio should change.

  We use the following parameters:

  1. The **time factor** determines the extent to which the time duration of an audio signal is stretched or compressed.

2.1 Tunning parameters

To avoid the audios becoming unintelligible after the data augmentation, we have examined the performance of the transcriptions according to the parameter values and, based on the results, we have determined the ranges that we will use with the audios to be processed. For this purpose, we use Whisper (Radford et al., 2022) which is a state-of-the-art speech recognition system from OpenAI.[2]

We used the **Word Error Rate (WER)** metric to evaluate transcription performance (Su et al., 1992). It measures the percentage of incorrectly recognized words in the transcribed text compared to the original spoken text. Word errors include the insertion, deletion, or substitution of a word in the text generated by the recognition system. A lower WER indicates better accuracy and a higher level of speech recognition performance. It is computed as follows:

$$WER = \frac{(S + D + I)}{N}$$

Where `S` is the number of word substitutions, `D` is the number of word deletions, `I` is the number of word insertions, and `N` is the total number of words in the reference transcription.

We employed a collection of audio files for each language to adjust the parameters[3]. To achieve this, we determined the WER metric using the transcriptions of the original audio files and the transcriptions of the audio files after applying ADA with modified parameters. In order to identify the acceptable range of parameter values that maintain speech intelligibility, a tolerance parameter was established. The tolerance parameter was set to 0.33.

Figure 3 illustrates the evolution of WER as a function of the parameters utilized for each technique, with three instances presented as examples. It helped us to ensure that the generated samples were of high quality and did not lead to significant performance degradation.

Table 1 shows the range of the parameter based on the previous evaluation:

It is important to highlight that while we obtained the range of parameters by analyzing the Word Error Rate (WER) for each technique, the results could be influenced by the transcription tool used (in this case, Whisper). The accuracy of any speech recognition system, including Whisper, is affected by a variety of factors, such as the quality of the audio input and the language of the transcription. It is crucial to consider this limitation when interpreting the results and to evaluate the performance of these techniques using other transcription tools and datasets to ensure their effectiveness and generalizability.

[2] https://github.com/openai/whisper

[3] We use the languages available in the Whisper model.

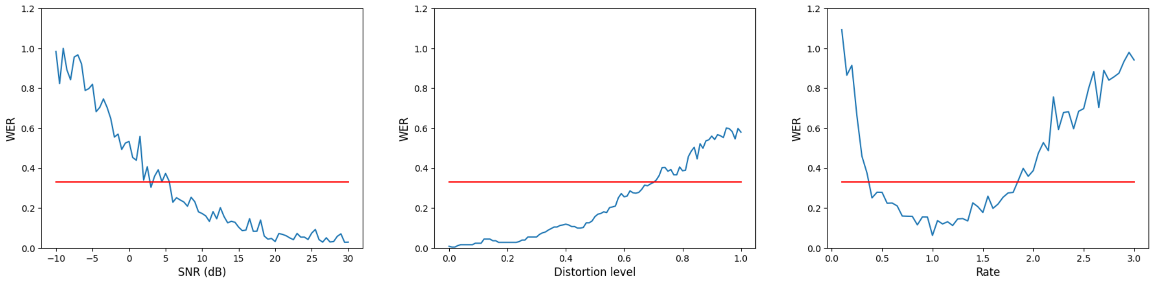


Figure 3: Three examples of transcription performance versus transformation parameters.

ADA technique	Parameter name	Range
Background noise	SNR (dB)	6.00 - 30.00
Gaussian noise	amplitude	0.01 - 0.025
Tanh Distortion	distortion level	0.00 - 0.70
Time stretch	time factor	0.40 - 1.80

Table 1: Parameters used in each ADA technique.

For the ADA process, we have applied a background noise, followed by a transformation that allows us to expand the original dataset many times. In our case, we increased the data 20 times.

The resulting audio files generated through the augmentation process must be realistic and representative of the original audio data. It is also possible to further increase the dataset size by incorporating additional transformations. It is crucial to ensure that the resulting audio files remain authentic and do not deviate significantly from the original audio data.

3 Methodology

Our proposed methodology aims to be applied in any language. For that reason, we first describe the workflow to generate the corpus using ADA. Then, in section 4, we explain how to apply this methodology in a real project using a data augmentation project for the languages of Spain. Our methodology is presented in 4.

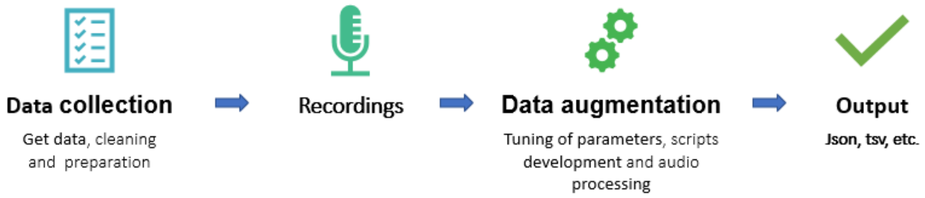

Data collection
Get data, cleaning and preparation
Recordings
Data augmentation
Tuning of parameters, scripts development and audio processing
Output
Json, tsv, etc.

Figure 4: Proposed methodology for carrying out a speech project with data augmentation.

As with any project, there is a planning phase to make crucial decisions useful for the subsequent phases. To ensure effective planning, it is essential to make these decisions at this stage rather than at a later phase. Among the most important decisions is the choice of language (as it will impact data collection), participant recruitment, and the recording phase.

It is necessary to determine the context of the data (such as text related to general news, social media, health, politics, and other relevant topics), the length of text segments to be used in recordings, the required number of participants, and their age range. It is also important to define the audio format, background noise restrictions, and the setting for recording sessions (e.g., in a studio or online). Lastly, it is necessary to define the ADA techniques.

For the **data collection** phase, the decisions made in the initial phase are implemented, which involves extracting the required amount of text segments for each language within the defined context. Also, if it is required, we normalize and prepare the text for the next phase.

In the **recording** stage, participants do record the text segments that have been assigned in their language. Proper follow-up and organization of the recording sessions are necessary to ensure the accurate documentation of relevant information. Every captured audio requires the storage of its file and duration, along with the accompanying text and demographic attributes of the speaker, such as age range and gender. Additionally, we must record the format of the recording, including sample rate, audio type, and duration. Organizing the process carefully guarantees that all relevant data is precisely documented and readily accessible for future use. Upon completion of the recording stage, it would be advantageous to have an automated process or human review that can verify the correctness of the audio recordings and the existence of the files.

During the **ADA**, the techniques defined in the initial phase should be applied to increase the diversity of the dataset and to avoid any patterns or predictability in the dataset. When incorporating background noise, such as street sounds, stadium noise, or subway sounds, it is advisable to divide it into smaller segments. In addition, it is important to carefully adjust the parameters of the techniques to avoid creating indecipherable sounds that could negatively affect the quality of the dataset. By randomly and systematically applying data augmentation techniques and carefully fine-tuning the parameters, the dataset can be diversified and of high quality.

To produce the desired output, a final file should be created containing the metadata of each audio file. This metadata must include the transcription, audio file name, speaker characteristics, and audio details.

The proposed methodology is intended to simplify the implementation of new projects. In the subsequent section, we offer a real project example to demonstrate that the methodology is well-suited for such initiatives.

4 Applying ADA

The methodology presented in section 3 provides a comprehensive guide for obtaining an extensive speech dataset with its transcriptions. In this section, we apply the methodology in a real speech augmentation project for Spanish languages.

We have followed our proposed methodology's sequence of steps, which consists of collecting data, conducting recording sessions, and carrying out the augmentation.

For practical purposes, we use the term **participants** to refer to the individuals who record the audio and **coordinator** to refer to the person who manages the recording process.

To begin with, we make several initial decisions that will be useful for each of the following phases.

For the **data collection** phase:

- We agreed on using 4 languages: Euskera, Galician, Asturian, and Catalan.
- We extracted text from online news media such as a segment of text in Catalan: *"Actualment tota la zona de conreu és edificada."*.

- Text segments should last between 5 and 30 seconds such as, similar to Radford et al. 2022, given that there are people who speak faster or slower, and segments can vary in length.

For the **recording process**[4]:

- Participants had to sign a consent form for the use of their voice (a draft is available in Appendix 1).
- Participants had to read the instructions for the recording (Appendix 2).
- We expected to recruit at least 50 participants. We also considered the balance in each age range (18-29, 30-49, and over 50) and gender.
- Recordings were conducted online.
- We requested the recordings in a noise-free environment, to be able to apply ADA. The recordings were in WAV, MP3 or MP4 format, with a sample rate of 8 kHz[5], which it should be achievable with most recording devices nowadays.

For the **DA processing**:

- We applied ADA techniques with the respective parameters shown in section 2. That is, for each audio file, we applied one background noise (randomly) followed by one of the three transformations (randomly).

4.1 Data collection

In this phase, we generated text segments for each selected language and participant. To achieve this, we used the Selenium tool[6], which allowed us to capture texts from different websites. We then used a segmentation tool called Segmenter [7] to cut the text into segments.

In our case, we did not apply any data cleaning or normalization since the texts were taken from news sites where the writing is generally quite good.

The complete process took around 5 days. It was challenging to find web pages in certain languages.

4.2 Recordings

In this phase, it was necessary to search for participants who can record in the selected languages. To achieve this, we posted advertisements in social organizations dedicated to language preservation, as well as on job portals. We were able to recruit 55 participants: 32 females and 23 males.

Before the recording process begins, it is essential that the participants have already signed the consent form. Subsequently, they should receive detailed instructions and recommendations on how to carry out the recordings (refer to Appendices 1 and 2 for guidance).

The coordinator is responsible for sending the text segments to participants for recording. Additionally, the coordinator must register the gender, age range, and language of each participant to include them in the metadata of each audio file. After recording the audio files,

[4] We used an in-house tool for recording sessions. Nevertheless, this is not a requirement, and anyone can manually reproduce this workflow

[5] If higher quality is needed, we recommend keeping the **WAV** format and increasing the sampling rate to at least 22 kHz.

[6] https://www.selenium.dev/documentation/

[7] https://github.com/diasks2/pragmatic_segmenter

participants send them all to the coordinator, ideally named with an identifier that enables the coordinator to match them with the corresponding text segment. Figure 5 shows this process.

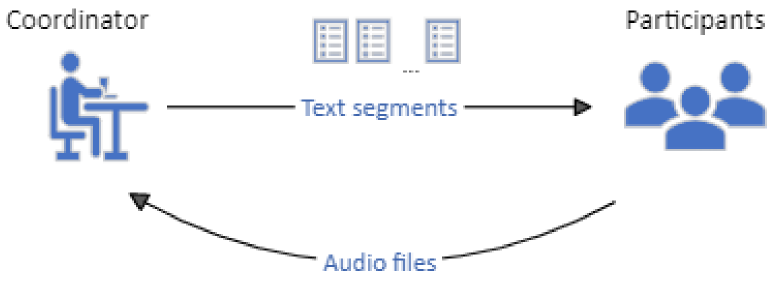


Figure 5: Coordinator assigns text segments to each participant, based on their assigned language. Participants complete the recordings and send audio files to the coordinator.

The entire process took about a month to complete and the results are shown in Table 2.

Language	**# audios**	**hrs**	**% female/male**
Euskera	12,231	37.6	60.4% / 39.6%
Catalan	20,277	56.7	57.6% / 42.4%
Asturian	8,459	25.6	63.4% / 36.6%
Galician	22,304	62.5	47.1% / 42.9%
Total	**63.271**	**182.5**	

Table 2: Final dataset

We could have made improvements such as adding a check process to validate that the recordings were made correctly. For example, incorporating existing transcription tools in these languages, such as Whisper (Radford et al., 2022) and comparing the original text with the audio transcription.

4.3 Audio Data Augmentation processing

In this subsection, we applied ADA to the recorded audio files. In particular, we add one random background noise to the audio. Then, we apply a random transformation (in our case, we use one transformation, but we could use more than one). Finally, we get the transformed audio and its new metadata file, which contains the original audio information and the applied transformation information. Figure 6 shows this pipeline.

4.4 Output

When delivering a file as output, it is important to ensure that it is in a format that can be easily understood and manipulated by machines. One such format that is commonly used for this purpose is JSON (JavaScript Object Notation). It is lightweight, easy to parse (by machines), platform-independent, human-readable, and easy to generate. In many programming languages, it is also useful for representing structured data.

Figure 7 shows an example of the metadata of one file (without applying ADA) and Figure 8 shows the metadata of one file after ADA. Delivering information in this particular format

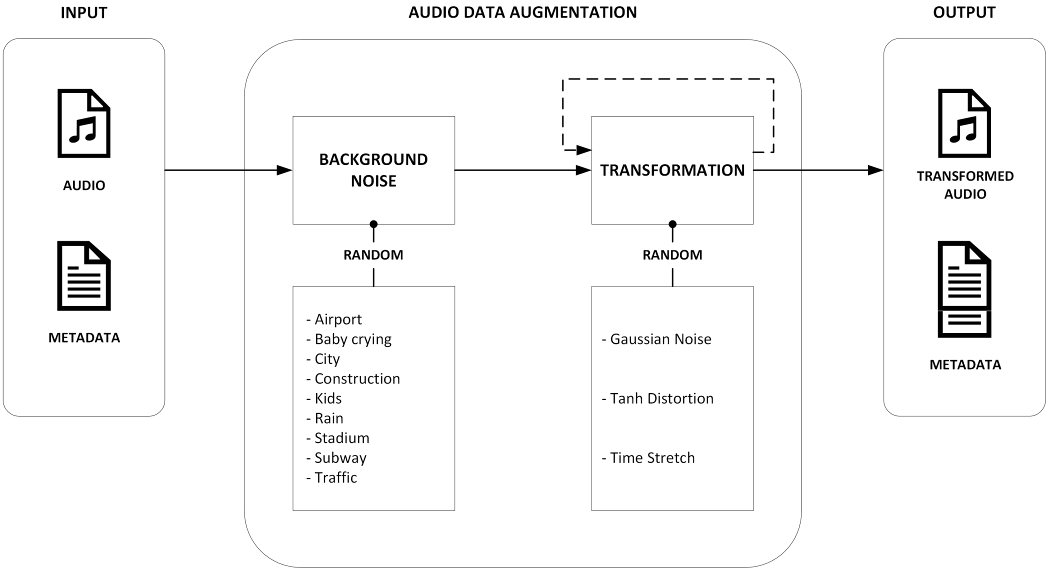
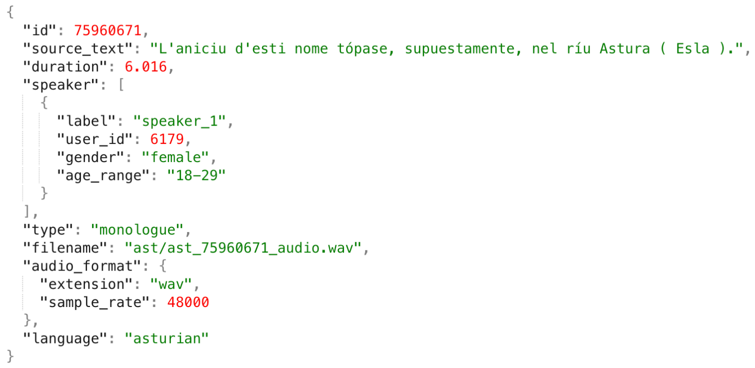


Figure 6: Pipeline of ADA. We could use more than one transformation.



```json
{
  "id": 75960671,
  "source_text": "L'aniciu d'esti nome tópase, supuestamente, nel ríu Astura ( Esla ).",
  "duration": 6.016,
  "speaker": [
    {
      "label": "speaker_1",
      "user_id": 6179,
      "gender": "female",
      "age_range": "18-29"
    }
  ],
  "type": "monologue",
  "filename": "ast/ast_75960671_audio.wav",
  "audio_format": {
    "extension": "wav",
    "sample_rate": 48000
  },
  "language": "asturian"
}
```

Figure 7: Metadata without applying ADA.

enables us to ensure its effortless parsing and manipulation by machines. As a result, integration with other software systems and automation of data processing tasks become more convenient.

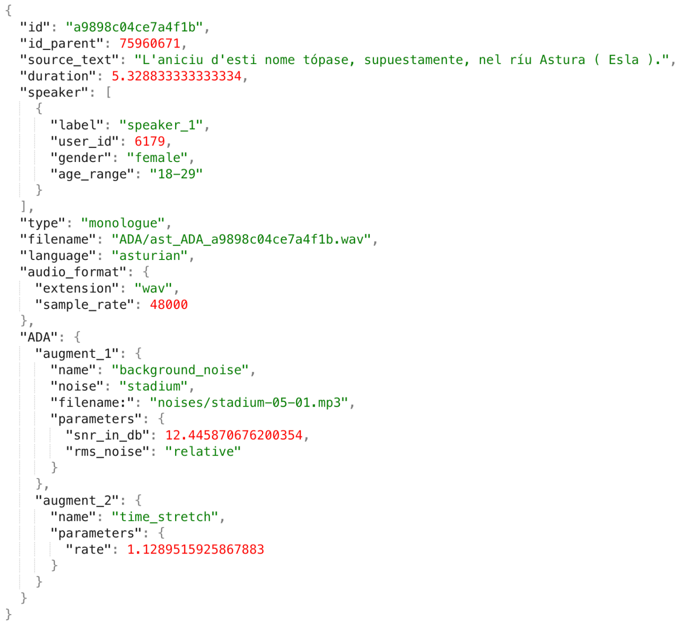
The dataset is available on Github[8].

5 Discussion

In this section, we provide an overview of the key highlights and challenges encountered during the development of our project. We also share some valuable suggestions to consider if you plan to undertake a speech project that involves data augmentation.

We discovered several data augmentation techniques that could potentially enhance the quality of our speech recognition model (Feng et al., 2021; Shorten et al., 2021). After careful

[8] https://github.com/Pangeamt/ele2-ada



```
{
  "id": "a9898c04ce7a4f1b",
  "id_parent": 75960671,
  "source_text": "L'aniciu d'esti nome tópase, supuestamente, nel ríu Astura ( Esla ).",
  "duration": 5.328833333333334,
  "speaker": [
    {
      "label": "speaker_1",
      "user_id": 6179,
      "gender": "female",
      "age_range": "18-29"
    }
  ],
  "type": "monologue",
  "filename": "ADA/ast_ADA_a9898c04ce7a4f1b.wav",
  "language": "asturian",
  "audio_format": {
    "extension": "wav",
    "sample_rate": 48000
  },
  "ADA": {
    "augment_1": {
      "name": "background_noise",
      "noise": "stadium",
      "filename:": "noises/stadium-05-01.mp3",
      "parameters": {
        "snr_in_db": 12.445870676200354,
        "rms_noise": "relative"
      }
    },
    "augment_2": {
      "name": "time_stretch",
      "parameters": {
        "rate": 1.1289515925867883
      }
    }
  }
}
```

Figure 8: Metadata after applying ADA.

consideration, we decided to use and adapt the **audiomentations** library, which offers a range of customizable techniques, flexibility, ease of use, and clear documentation. We found this to be a practical choice since it utilizes well-known libraries in the background such as **librosa**[9] for processing audio files.

While exploring other approaches, we also investigated the use of deep learning techniques (Shorten and Khoshgoftaar, 2019), We considered that these methods were not necessary for the current project but they could be explored in future studies to improve the performance of the model potentially.

While we were able to record the necessary data, the process was not without its challenges. Here are some of the main **challenges** we encountered during the recording process and how we overcame them:

- Initially, we defined the age ranges as <18, 18-29, 30-59, and 60>. However, due to legal restrictions, we had to exclude those under 18 from our study. Furthermore, we faced a challenge in recruiting individuals over the age of 60 to participate in the recording. This may be because older adults are often less familiar with technology and may not have access to recording devices. As a result, we could only recruit 3 participants over the age of 60, out of the 4 languages we were working with. To address this imbalance, we reorganized our age ranges to 18-29, 30-49, and 50>. This allowed us to achieve a more balanced and representative dataset.

- Despite advertising in various media channels to attract participants, we encountered a challenge in finding participants who could record in some languages, as relatively small communities speak these languages in Spain. Consequently, we had to expand our language selection to increase our pool of available participants. For instance, we planned to record in the Aranese and Aragonese languages, but we faced challenges in finding participants and could find only two people in total. As a solution, we re-

[9] https://librosa.org/

placed these with the Euskera and Asturian languages, which helped us to attract more participants.

- It was challenging to achieve an appropriate gender and age balance. In future projects, it is crucial to recognize the significance of promoting diversity and gender equality in the dataset, and we also consider this aspect as we continue to refine our research approach.

Along with the positive aspects and challenges we encountered, here are some **suggestions** to consider for a future speech recording project that involves ADA.

- When creating background noise audio files, it is important to maintain a balance of each type. For instance, if we decided to use 20 instances of street noise, the remaining background noises should also be set to 20. It is advisable to have a significant variety of each type of noise to prevent a speech recognition model from learning to recognize the background noise during training.
- Programming skills are crucial during the project, as you need to operate and create `JSON` files, collect and organize data for participants, edit data augmentation functions, and perform other technical tasks. Therefore, it is essential to have programming expertise.
- Regarding the recording tool, it can be any device that is capable of recording audio such as a phone, computer, or any other recording device. For our project, we used an in-house tool that enables us to efficiently manage and organize the audio recordings for participants. However, it is not a mandatory requirement to have a specialized tool for carrying out the project.
- While the audio format can be a crucial factor in the recording process, it is advisable to ensure that participants are capable of meeting the necessary format requirements if it is deemed important. In our particular case, we do not consider it to be a mandatory requirement. However, if the quality is a concern, it is recommended to use a high-quality format, such as a sample rate of 16 kHz or higher, and a lossless audio file format such as `WAV`.
- The versatility of the `JSON` file format makes it advantageous for data manipulation. In the Python programming language, `JSON` files can be easily opened, filtered, searched, and analyzed using graphical tools. Additionally, `JSON` files are useful for conducting data analysis. To facilitate this process, we recommend using the Pandas[10] library which allows for seamless manipulation of these files.
- Our project involved collecting $63,271$ voice recordings along with their transcriptions and associated metadata.While this is a significant increase, it is still inadequate for training a Machine Learning model. To ensure optimal performance, it is crucial to have a diverse and varied dataset. This includes incorporating a wider range of background noises and applying additional augmentation techniques (and combining them). By doing so, we can mitigate any potential biases or imbalances that Machine Learning models are susceptible to.
- In our case, we were able to increase the size of our dataset by a factor of 20. Nonetheless, considering that we have 9 different background noises and 3 ADA techniques, we could have achieved more. This implies that we have the potential to increase our dataset size by a factor of 27. And also, if we were to incorporate additional ADA techniques, we could potentially expand our dataset even further.

[10] https://pandas.pydata.org/

- Although we managed to acquire numerous voice recordings, we lacked the ability to confirm whether each of the captured audios matched the original text precisely. Consequently, it would be beneficial to include a quality assurance step after the recordings. For the same reason, some of the audios of the final dataset may not match completely with the associated text segment.

Despite the challenges we initially faced in understanding data augmentation techniques and organizing our work plan, our team gained extensive knowledge through this project. As a result, we have plans to continue exploring related projects at the company level.

Furthermore, we are continuously exploring new software to enhance audio augmentation capabilities. In particular, we are looking forward to seeing the potential of GPT-4[11], a large multimodal model that we expect to be released in the future. While we have not had the opportunity to test GPT-4 during this project, we are excited about its potential for data augmentation. We think that it would add value to the project.

6 Summary and conclusions

Recognition systems are increasingly present in our daily lives, from virtual assistants to smart home devices. Unfortunately, EU language datasets for speech technology are lacking today. However, data augmentation techniques can play a crucial role in developing a large, realistic, and more comprehensive dataset.

We presented a methodology for carrying out a speech project that enables the extension of a dataset up to 20 times using advanced data augmentation techniques. To demonstrate the effectiveness of our approach, we apply the methodology in a real project using four languages of Spain. We followed the five-phase process outlined in the methodology to ensure successful project completion, from initial planning to final delivery.

This proposal aligns with the strategic agenda of ELE by providing a comprehensive dataset comprising realistic audio environments and corresponding transcriptions. This dataset can be highly valuable in training Machine Learning models for speech recognition, and it includes metadata such as the gender and age range of each speech voice. The guidelines presented in this report offer a structured approach that can be applied to other EU languages and different scenarios. We hope that our proposal will inspire the speech technology community to develop more extensive speech datasets in other EU languages to preserve them for the future.

References

ELE Consortium. Digital Language Equality in Europe by 2030: Strategic Agenda and Roadmap, 2022.

Steven Y. Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. A survey of data augmentation approaches for nlp, 2021.

Lucas Ferreira-Paiva, Elizabeth Alfaro-Espinoza, Vinicius M Almeida, Leonardo B Felix, and Rodolpho VA Neves. A survey of data augmentation for audio classification. 2021.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022.

Justin Salamon and Juan Pablo Bello. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal processing letters*, 24(3):279–283, 2017.

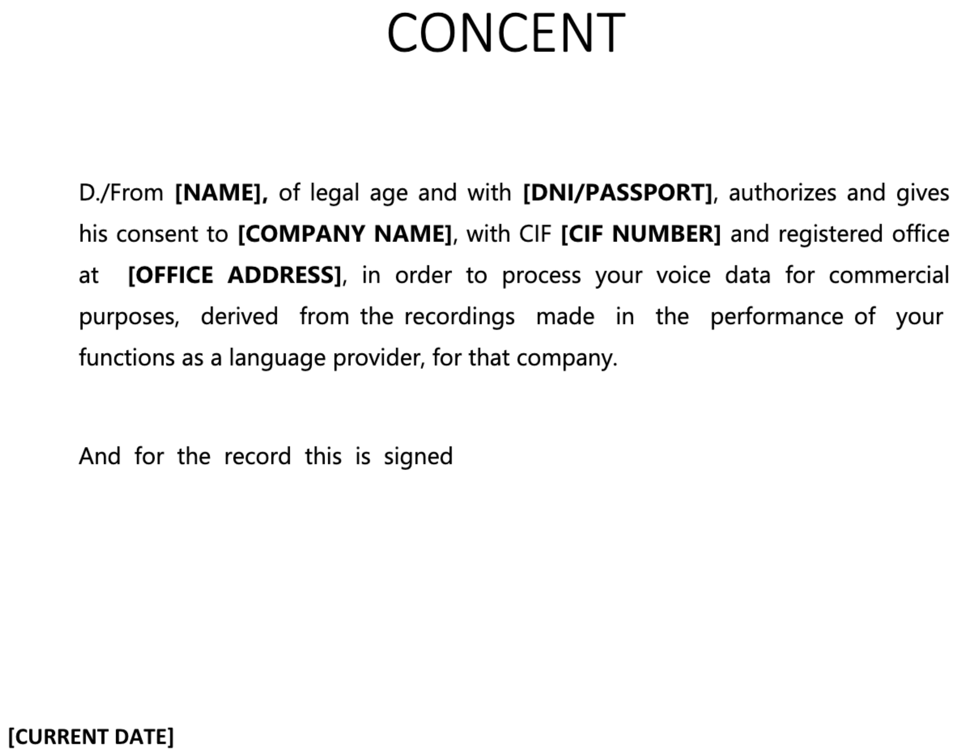
[11] https://openai.com/research/gpt-4

Connor Shorten and Taghi M Khoshgoftaar. A survey on image data augmentation for deep learning. *Journal of big data*, 6(1):1–48, 2019.

Connor Shorten, Taghi M Khoshgoftaar, and Borko Furht. Text data augmentation for deep learning. *Journal of big Data*, 8:1–34, 2021.

Keh-Yih Su, Ming-Wen Wu, and Jing-Shin Chang. A new quantitative quality measure for machine translation systems. In *Proceedings of the 14th Conference on Computational Linguistics - Volume 2*, COLING '92, page 433–439, USA, 1992. Association for Computational Linguistics. doi: 10.3115/992133.992137. URL https://doi.org/10.3115/992133.992137.

Appendix 1

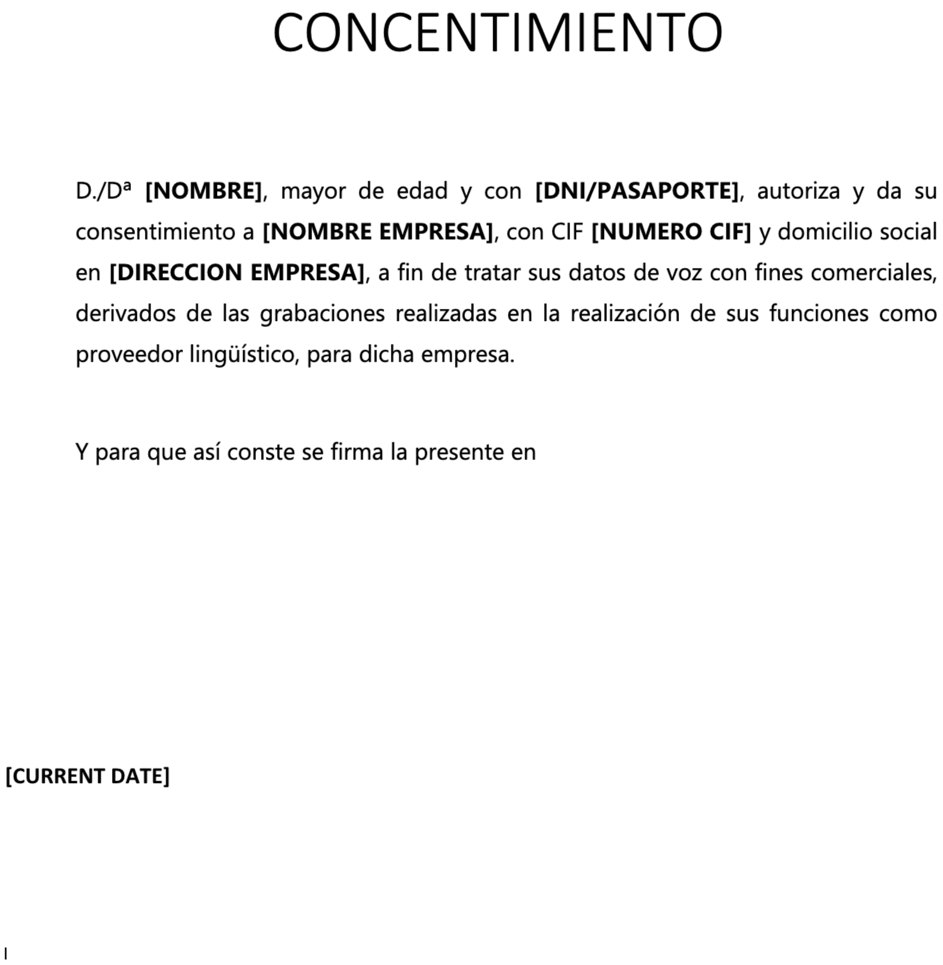


CONCENT

D./From **[NAME],** of legal age and with **[DNI/PASSPORT]**, authorizes and gives his consent to **[COMPANY NAME]**, with CIF **[CIF NUMBER]** and registered office at **[OFFICE ADDRESS]**, in order to process your voice data for commercial purposes, derived from the recordings made in the performance of your functions as a language provider, for that company.

And for the record this is signed

[CURRENT DATE]



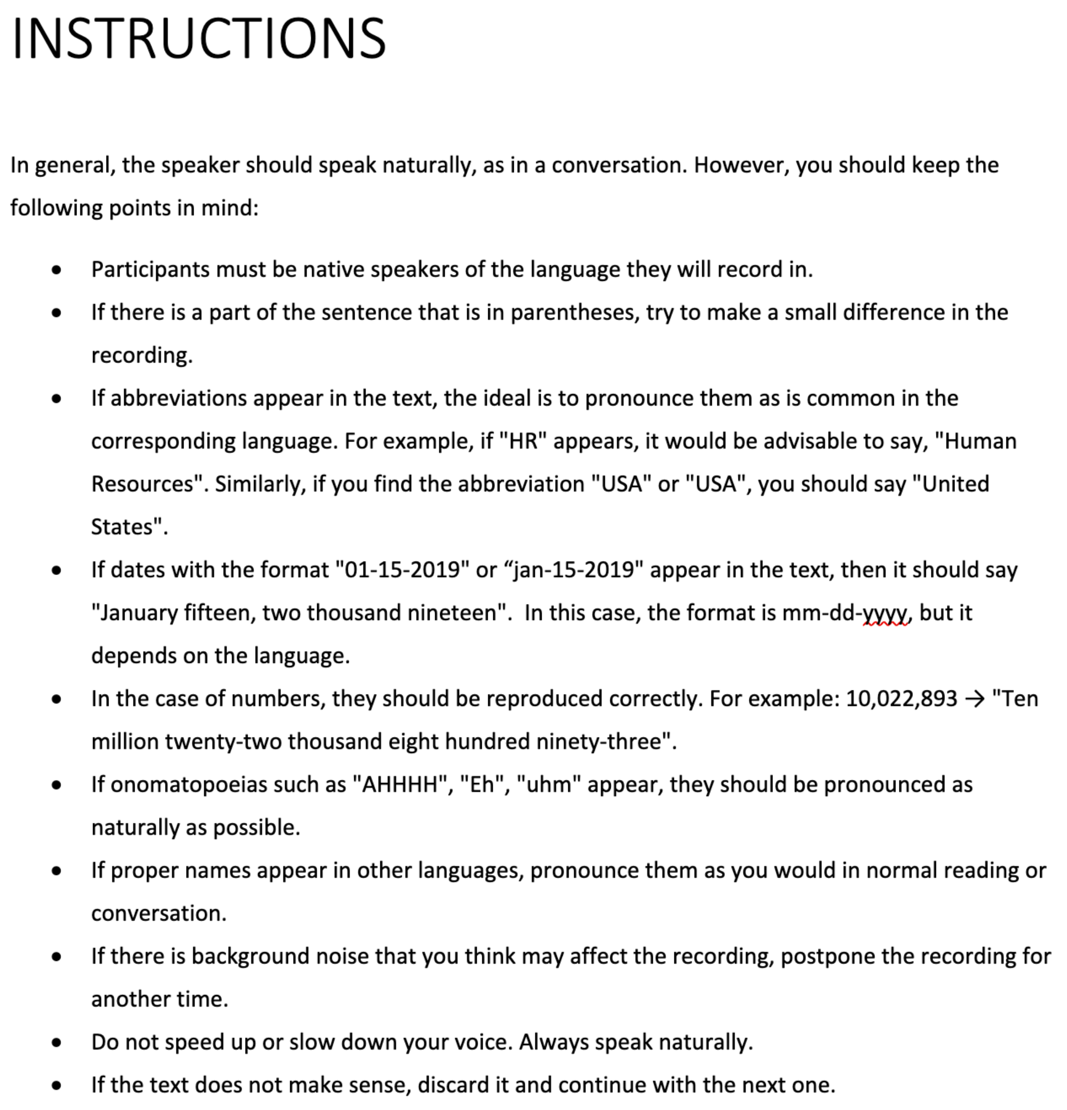
CONCENTIMIENTO

D./Dª **[NOMBRE]**, mayor de edad y con **[DNI/PASAPORTE]**, autoriza y da su consentimiento a **[NOMBRE EMPRESA]**, con CIF **[NUMERO CIF]** y domicilio social en **[DIRECCION EMPRESA]**, a fin de tratar sus datos de voz con fines comerciales, derivados de las grabaciones realizadas en la realización de sus funciones como proveedor lingüístico, para dicha empresa.

Y para que así conste se firma la presente en

[CURRENT DATE]

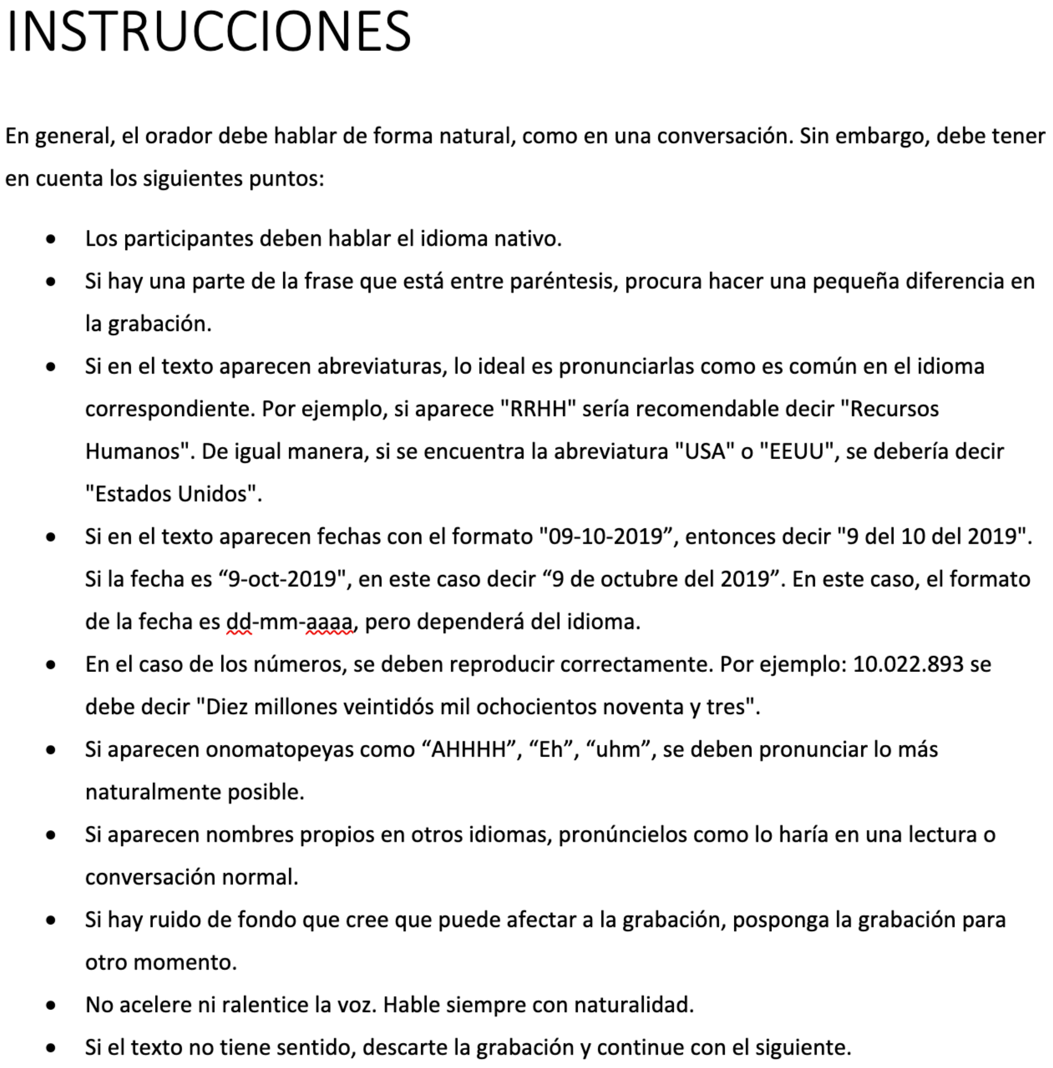
Appendix 2



INSTRUCTIONS

In general, the speaker should speak naturally, as in a conversation. However, you should keep the following points in mind:

- Participants must be native speakers of the language they will record in.
- If there is a part of the sentence that is in parentheses, try to make a small difference in the recording.
- If abbreviations appear in the text, the ideal is to pronounce them as is common in the corresponding language. For example, if "HR" appears, it would be advisable to say, "Human Resources". Similarly, if you find the abbreviation "USA" or "USA", you should say "United States".
- If dates with the format "01-15-2019" or “jan-15-2019" appear in the text, then it should say "January fifteen, two thousand nineteen". In this case, the format is mm-dd-yyyy, but it depends on the language.
- In the case of numbers, they should be reproduced correctly. For example: 10,022,893 → "Ten million twenty-two thousand eight hundred ninety-three".
- If onomatopoeias such as "AHHHH", "Eh", "uhm" appear, they should be pronounced as naturally as possible.
- If proper names appear in other languages, pronounce them as you would in normal reading or conversation.
- If there is background noise that you think may affect the recording, postpone the recording for another time.
- Do not speed up or slow down your voice. Always speak naturally.
- If the text does not make sense, discard it and continue with the next one.



INSTRUCCIONES

En general, el orador debe hablar de forma natural, como en una conversación. Sin embargo, debe tener en cuenta los siguientes puntos:

- Los participantes deben hablar el idioma nativo.
- Si hay una parte de la frase que está entre paréntesis, procura hacer una pequeña diferencia en la grabación.
- Si en el texto aparecen abreviaturas, lo ideal es pronunciarlas como es común en el idioma correspondiente. Por ejemplo, si aparece "RRHH" sería recomendable decir "Recursos Humanos". De igual manera, si se encuentra la abreviatura "USA" o "EEUU", se debería decir "Estados Unidos".
- Si en el texto aparecen fechas con el formato "09-10-2019", entonces decir "9 del 10 del 2019". Si la fecha es "9-oct-2019", en este caso decir "9 de octubre del 2019". En este caso, el formato de la fecha es dd-mm-aaaa, pero dependerá del idioma.
- En el caso de los números, se deben reproducir correctamente. Por ejemplo: 10.022.893 se debe decir "Diez millones veintidós mil ochocientos noventa y tres".
- Si aparecen onomatopeyas como "AHHHH", "Eh", "uhm", se deben pronunciar lo más naturalmente posible.
- Si aparecen nombres propios en otros idiomas, pronúncielos como lo haría en una lectura o conversación normal.
- Si hay ruido de fondo que cree que puede afectar a la grabación, posponga la grabación para otro momento.
- No acelere ni ralentice la voz. Hable siempre con naturalidad.
- Si el texto no tiene sentido, descarte la grabación y continue con el siguiente.