VoxCeleb-PT – a dataset for a speech processing course

John Mendonça, Isabel Trancoso

INESC-ID/Instituto Superior Técnico, University of Lisbon, Portugal
{john.mendonca, isabel.trancoso}@tecnico.ulisboa.pt

Abstract

This paper introduces VoxCeleb-PT, a small dataset of voices of Portuguese celebrities that can be used as a language-specific extension of the widely used VoxCeleb corpus. Besides introducing the corpus, we also describe three lab assignments where it was used in a one-semester speech processing course: age regression, speaker verification and speech recognition, hoping to highlight the relevance of this dataset as a pedagogical tool. Additionally, this paper confirms the overall limitations of current systems when evaluated in different languages and acoustic conditions: we found an overall degradation of performance on all of the proposed tasks.

Index Terms: speech datasets, speech processing, speaker recognition, paralinguistics, speech recognition

1. Introduction

Teaching a one-semester MSc course on speech processing at the Engineering school of the University of Lisbon has always been a huge challenge because of the breadth of the topic and the diversity of backgrounds. These can range from more signal-processing oriented to more computer science oriented students, but occasionally including external students from other faculties with a much more diverse background (e.g. classical linguistics). Teaching the same course in the deep learning era is a much bigger challenge, given the huge technological advances we witnessed during the last 5 years [1].

The goal of this paper is not to describe the choice of the curriculum, but to introduce a small dataset – VoxCeleb-PT. This dataset served as a basis for some of the assignments of the latest edition of this course. VoxCeleb-PT is intended to be an extension of the widely used VoxCeleb corpus [2, 3]. The original corpus includes recordings of 7,363 speakers of multiple ethnicities, accents, and age groups. It is composed of short clips taken from interviews of celebrities uploaded to YouTube and is composed of two parts, VoxCeleb 1 and 2, both subdivided into dev and test.

Our goal in asking the students of the course to contribute to the collection of interviews from Portuguese celebrities was twofold: on one hand, we wanted the students to get familiar with the data collection process, including metadata labeling as a first course assignment. On the other hand, we wanted them to use this dataset in several speech processing tasks, in subsequent assignments.

As in previous years, we tried to include a paralinguistic task (for instance, detection of cold, sleepiness, depression, etc.) as one of the assignments. Typically, such tasks involve the use of a conventional machine learning method such as SVM (Support Vector Machine), due to the relatively small datasets one finds in paralinguistic tasks that are not adequate for training very deep neural networks. The selected paralinguistic task for this year was age regression, given the existence of a widely distributed training corpus and the fact that age information was available for all the celebrities, enabling the use of VoxCeleb-PT as a test set. This task could also allow students to investigate potential mismatch problems when using the small corpus for testing systems trained with a different language domain, in tasks that are supposed to be relatively language independent such as age regression.

Due to the recent advances achieved by pre-trained models in several speech and language tasks, we also tried to include a speaker recognition assignment which could leverage such speaker representations, and where VoxCeleb-PT could serve as a test set of embeddings trained on VoxCeleb 1 and 2.

The last assignment using the recently collected dataset was just an assessment task: we wanted the students to assess the speech recognition system used on YouTube for European Portuguese.

The remainder of this paper is organised as follows: Section 2 describes the dataset, together with the guidelines given to the students for its collection; Section 3 details the three assignments using this dataset, including where appropriate the baseline experiments that the students were asked to improve, and the results they achieved. Finally, Section 4 presents our conclusions and topics for future work.

2. Dataset

In this first assignment, students were asked to collect an in-the-wild dataset with interviews of Portuguese celebrities, playing the role of an expert annotator. The deliverable for this individual assignment was a structured data directory with:

- 20-25 minutes of spontaneous speech from the selected celebrity (after removing speech from others, noise and/or music). This constraint likely requires the need to include more than one video.
- Transcriptions from the subtitle frames in which the celebrity is speaking.
- Metadata, including gender and age (at the time of the video).

The celebrity should be an adult Portuguese person with recognizable internet presence, including those who passed away. The choice of the celebrities was totally up to the students who initially committed to collecting the corresponding data on a first come first served basis, in order to avoid repeated celebrities. Actors, TV presenters, politicians and football celebrities were the most popular. An effort was made to promote the selection of more female celebrities, as the initial set of commitments was heavily biased towards male celebrities.

The lab guide used a football celebrity as example (as shown in Figure 1), starting by what could be a convenient query on YouTube (e.g. Cristiano Ronaldo entrevista (interview)), and suggesting the choice of long interviews where the interviewee has plenty of time to talk. Next, the lab guide recommended how to download the automatic transcription (in...
At this stage in the course, students were recommended to do manual diarization of the interviews instead of using an automatic tool, in order to assign each caption to either the celebrity or the interviewer. Voice activity detection was also implicitly accomplished in this process. Some pre-processing scripts were recommended to ensure that all captions were in the same format. Hints on how to obtain timestamps with millisecond information were also given (e.g. right-clicking video > Stats for Nerds > Mystery Text > t:s.mm). The next recommended step was video segmentation according to the timestamps, using ffmpeg\(^2\) in a Python for loop.

Table 1 shows the demographic characteristics of the collected dataset of 51 celebrities. Altogether, VoxCeleb-PT contains 26,663 automatically transcribed utterances (.wav 16kHz, pcm_s16le). The total duration is 17:55:14, with an average of 20 min/spk, and utterance duration 2-5s. Contrarily to specifications, three of the selected male celebrities (in the age ranges 20-29, 60-69, 40-50) were recorded doing formal speeches instead of interviews.

<table>
<thead>
<tr>
<th>Age range</th>
<th>F</th>
<th>M</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>&gt;670</td>
<td>2</td>
<td>3</td>
<td>5</td>
</tr>
<tr>
<td>60-69</td>
<td>2</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>50-59</td>
<td>2</td>
<td>5</td>
<td>7</td>
</tr>
<tr>
<td>40-49</td>
<td>6</td>
<td>8</td>
<td>14</td>
</tr>
<tr>
<td>30-39</td>
<td>7</td>
<td>5</td>
<td>12</td>
</tr>
<tr>
<td>20-29</td>
<td>4</td>
<td>3</td>
<td>7</td>
</tr>
<tr>
<td>Total</td>
<td>23</td>
<td>28</td>
<td>51</td>
</tr>
</tbody>
</table>

Table 1: Age and gender statistics on VoxCeleb-PT.

The dataset can be obtained in its original form and with both development (train/val) and test sets, all containing the full speaker cohort. All sets contain speaker id, age, gender and manually corrected transcriptions. As such, VoxCeleb-PT should prove useful for a variety of tasks, namely ASR, Speaker Verification and Age/Gender Recognition.

VoxCeleb-PT is available for download at [https://www.hlt.inesc-id.pt/w/VoxCeleb-PT](https://www.hlt.inesc-id.pt/w/VoxCeleb-PT).

### 3. Course assignments

This section describes the three assignments that were developed using the recently collected dataset. Although the dataset collection assignment was done individually, the remaining assignments were performed by groups of two students (24 groups).

#### 3.1. Age regression

In this assignment, the students were asked to develop their own machine learning models for the task of age regression. Model development was conducted on the INTERSPEECH 2010 Paralinguistic Challenge [4] Gender Subchallenge dataset, denoted aGender [5]. aGender consists of 47 hours of speech in 65,364 single utterances of 954 German speakers in a mobile phone environment. The utterances were stored on the application server as 8 bit, 8 kHz, A-law. The dataset was randomly partitioned according to speakers, resulting in a roughly 40% 30% Train/Develop/Test split. For evaluation, the Mean Absolute Error MAE (arithmetic average of the absolute errors \( |y_i - \hat{y}_i| = |y_i - x_i| \), where \( y_i \) is the prediction and \( x_i \) the true value) was used.

#### 3.1.1. Baseline

Research on age regression (and age classification) from speech has largely remained stale, mainly due to limitations regarding the existing gender datasets (aGender, SRE 2008 [6] and 2010 [7]), all of which are controlled telephony speech. Most works on age regression base themselves on SRE data and leverage speaker representation of utterances: a MLP with a single hidden layer that receives as input MFCC-based 400-dimensional i-vectors achieves 5.49 MAE for male speakers and 6.35 MAE for females [8]; phonetically-aware i-vector extractor paired with LDA results in a 4.7 MAE [9]; a fusion of i-vector and x-vector embeddings with LDA achieve 5.20 overall MAE [10]. More recently, an annotation effort has been conducted to provide speaker age labels for the VoxCeleb dataset [11]. The authors report a MAE of 9.443 using using i-vector features with ridge regression.

In order to simplify the process for students, two simple baselines were provided, which they could improve upon: a Support Vector Regressor (SVR) and a Neural Network (NN). In both cases, the Geneva Minimalistic Acoustic Parameter Set (GeMAPS) was used as the input features, but students had the liberty to experiment with other feature sets.

The SVR baseline consisted of a Linear SVR trained with scikit-learn toolkit. All parameters were left to their defaults, leaving students the opportunity to outperform the baseline with hyperparameter tuning. The NN baseline consisted of a Single-Layer Perceptron with a hidden size of 32. A ReLU was used as non-linearity. The network was trained using stochastic gradient descent using a learning rate of 0.003 for 100 epoch, and the best performing model on the development set was selected for testing.
3.1.2. Experimental Setup

Given the small size of VoxCeleb-PT, it was only used as an out-of-domain test set. It is important to mention the domain mismatch between aGender and VoxCeleb-PT, namely in terms of language (German vs Portuguese) and channel conditions (8kHz vs 16kHz). This opened an opportunity to quiz the students regarding domain mismatching, and gauge their expectations on results obtained on out-of-domain datasets.

3.1.3. Results

The baseline results for this experiment, together with the best performing model submitted by the students are presented in Table 2. Models are evaluated using MAE, in line with previous work.

All but two groups surpassed the baseline. The best performing model consists of a Feed Forward Network with 3 Linear Layers paired with a ReLU non-linearity, Batch Normalization and Dropout. The model was trained for 250 epochs with a Learning Rate of 0.01 and a batch size of 100. L2 norm decay was set to 0.01, and dropout to $p = 0.5$.

<table>
<thead>
<tr>
<th>Model</th>
<th>Dev MAE</th>
<th>Test MAE</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear SVR baseline</td>
<td>17.19</td>
<td>16.92</td>
</tr>
<tr>
<td>NN baseline</td>
<td>16.14</td>
<td>19.40</td>
</tr>
<tr>
<td>3-Layer NN</td>
<td>15.88</td>
<td>11.97</td>
</tr>
</tbody>
</table>

Table 2: Age regression results. Dev MAE denotes results obtained on the development set of aGender. Test MAE denotes results on the test set of VoxCeleb-PT.

An inspection of the largest deviations shows that they frequently occur for all the tested machine learning methods, which points to errors in the computation of the feature sets, namely pitch. In fact, the largest deviation occurred for a segment with a filled pause uttered by a creaky voice.

3.2. Speaker recognition

The main goal of the speaker recognition task was to acquaint students with the typical speaker recognition pipeline and associated concepts, namely: (1) enrollment; (2) verification; and (3) identification. Students were asked to perform speaker verification and identification on the VoxCeleb-PT dataset, using state of the art approaches.

3.2.1. Baseline

X-vectors, first presented by [12], were developed as an improvement of the i-vector system [13], replacing Joint Factor Analysis of Gaussian Mixture Model supervectors by embeddings extracted from a feedforward DNN. The network of the x-vector system is divided into two different levels: the frame level uses a time delay architecture (TDNN) that functions on speech frames, offering temporal context [14]; the segment level is connected to the frame level using a statistics pooling layer. This pooling layer calculates the mean and standard deviation from the aggregate output of the final frame level. This pooling procedure compiles information from the entire segment to subsequent layers. The training of the DNN is conducted using multi-class cross-entropy objective.

A more recent extractor is the so-called ECAPA-TDNN embeddings, introduced by [15]. While their architecture is largely based on the original x-vector, several improvements are incorporated. Firstly, the initial frame layers are restructured into one-dimensional Res2Net modules with impactful skip connections. Similarly to SE-ResNet [16], Squeeze-and-Excitation blocks (SE) are introduced in these modules to explicitly model channel inter-dependencies. The SE block expands the temporal context of the frame layer by re-scaling the channels according to global properties of the recording. Secondly, information is aggregated and features are propagated to different hierarchical levels. Finally, the statistics pooling module is improved with channel-dependent frame attention. This enables the network to focus on different subsets of frames during each of the channel’s statistics estimation. The network is trained by optimising the AAM-softmax [17] loss on the speaker identities in the training corpus. This loss directly optimises the cosine distance between the speaker embeddings.

The SpeechBrain toolkit [18] offers both embedding extractors trained on VoxCeleb+1+2. The x-vector embedding extraction followed the Kaldi Speech Recognition Toolkit [19] recipe for VoxCeleb [3]. Training was conducted on the dev portion of VoxCeleb1, plus all of VoxCeleb2, augmented with reverberation and music, babble and noise from the MUSAN corpus [20]. The features were 30-dimensional MFCCs obtained every 10ms with a frame length of 25ms, mean-normalised over a sliding window of up to 3 seconds. An energy-based Voice Activity Detection module filtered out non-speech frames. X-vectors were extracted from the last layers of the pre-trained DNN model (before the softmax layer), outputting 512-dimensional embeddings.

For the ECAPA-TDNN embeddings, and similar to the x-vector training, RIRs2 and MUSAN were used for data-augmentation purposes. The input features are 80-dimensional MFCCs from a 25 ms window, with a 10 ms frame shift.

A Probabilistic Linear Discriminant Analysis (PLDA) model was used to score pairs of x-vectors when performing verification [21]. For the ECAPA-TDNN embeddings, the cosine score was used.

The results obtained with the two types of embeddings with the original VoxCeleb 1 test set show the recent progress in the state of the art. For the Verification task, the x-vector+PLDA scoring approach achieves 3.2% EER and the ECAPA-TDNN+Cosine Scoring approach achieves 0.69%.

3.2.2. Experimental Setup

Unlike other tasks, students were asked to follow concrete steps, being only required to write partial code to achieve the solution. For both Verification and Identification subtasks, students had to extract embeddings from the training set of VoxCeleb-PT and calculate the mean speaker embedding. For verification, they had to define a trial file, where they compared every speaker embedding with all utterances from the open eval set. For the Identification set, students were asked to provide speaker labels to the closed test set by selecting the highest score among speaker embedding comparisons.

3.2.3. Results

Results are reported using typical speaker recognition metrics: Equal Error Rate (EER) and minDCF (minimum Decision Cost Function) at $p = 0.01$ and equal costs for the verification task; accuracy (%) for the identification task.

For the verification task, the X-vector+PLDA scoring approach achieves 8.5% EER and a minDCF of 0.009, whereas the ECAPA-TDNN+Cosine Scoring approach achieves 1.9% EER and a minDCF of 0.001. For the identification task, we only
requested students to report results using the best performing model (ECAPA-TDNN). An accuracy of 98% was achieved for this task.

Overall, we note some performance degradation for the VoxCeleb-PT dataset. Speech utterances from a given speaker contain information related to the acoustic environment, transmission channel, speaker traits (accent, stress, speaking style, etc.) and spoken language. All this information can be thought of as different dimensions of the speaker acoustic space. Most of the previous research on train-test mismatch compensation for speaker recognition focused on acoustic conditions (namely acoustic environment and transmission channel, but also stress, speaking style, etc.). Language mismatch has received less attention [22], but it may be the cause for the observed degradation.

3.3. Speech recognition

The goal of this assignment was for students to assess the speech recognition system used by YouTube for European Portuguese, simultaneously learning how to apply the metric (WER - Word Error Rate), and to identify the main sources of error.

3.3.1. Experimental Setup

The assignment was performed in groups of two students, hoping to foster discussion between them. The assignment consisted in selecting a segment of video from one of the celebrities in the group including around 50 words. The deliverable for this part of the lab was a file which included the original transcription provided by YouTube, the manually corrected transcription, and the manually computed WER for this segment.

3.3.2. Results

The average WER was 34.75% with a large standard deviation (18.14%). The use of jargon was a common source of error (e.g., *bué* instead of *muito* - very). Foreign words were another source, although not so frequent. Disfluencies also had impact in the WER, but unlike older system, much progress was observed in the transcription of disfluent segments which are frequently common in spontaneous speech.

Another source of error was the presence (in the automatic transcriptions) of expressions and grammatical constructions which are more frequent in Brazilian Portuguese (BP) than European Portuguese (EP). The use of clitics is, in fact, one of the distinguishing factors in the grammars of BP and EP, being much more frequent in the latter (e.g., *partiu meu coração (BP)* vs. *partiu-me o coração (EP)*) (broke my heart)). The frequent errors in words with clitics most probably derive from the fact that the training dataset includes a much larger quantity of BP data. Vowel reduction including quality change, de-voicing and deletion, is especially important for EP. This is because it is one of the features that distinguishes it from BP and makes it more difficult to learn for a foreign speaker. As a result of vowel deletion, rather complex consonant clusters can be formed within words and across word boundaries. In fact, many ASR errors were found around sequences of obstruents formed across word boundaries, very frequent due to plural words with */s*/ in word final position. (Example: a sequence such as *todos estes meios que nos dão para evoluir* was recognized as *todos os mexicanos não podem evoluir*).

The speaker with the highest WER (above 80%) used many jargon expressions, had a very high speaking rate, a very high pitch, and described an emotional moment. Contrasting with the informality of this example, other examples of high WER occurred for the initial part of formal speeches, namely when addressing the high dignitaries in the audience with honorific titles (e.g., *Excelentíssimo - Your Excellency*). This type of material is probably under-represented in the training data.

Most of these factors were pointed out by students in their reports, in particular leading to the general perception that the ASR system was not tuned to European Portuguese.

4. Conclusions

This paper introduced VoxCeleb-PT, a small dataset that can be used as a language specific extension of the widely used VoxCeleb corpus for European Portuguese. We tried to highlight its relevance as a pedagogical tool by describing the lab assignments on age regression, speaker verification and speech recognition that were done on the basis of this dataset, including the results that can be used as baseline for this corpus. Although there were several other lab assignments in this course, the ones that involved VoxCeleb-PT were very motivating and the students were really engaged, including ERASMUS students.

Given the recent progress with few-shot and zero-shot multi-speaker Text-to-Speech models [23, 24, 25], a corpus such as VoxCeleb-PT could also be used to generate voices of celebrities, using just a few seconds of their voices. This was in fact done on request for the Portuguese TV station RTP in order to raise awareness to deep fakes. Additionally, it alerted the public to the potential misuse of current speech technologies, both in terms of impersonation and spoofing speaker verification systems.

5. Acknowledgements

We would like to thank all students of the Summer 2021 Speech Processing Course, which helped collect this dataset. This work was supported by Portuguese national funds through Fundação para a Ciência e a Tecnologia (FCT), with references UIDB/50021/2020, PRT/BID/152198/2021 and the P2020 program MAIA (LISBOA-01-0247-FEDER-045909).

6. References


