statistics

Huqariq: A Multilingual Speech Corpus of Native Languages of Peru for Speech Recognition

Rodolfo Zevallos, Luis Camacho, Nelsi Melgarejo

Universitat Pompeu Fabra, Pontificia Universidad Católica del Perú

Introduction

The Huqariq project responds to the endangerment currently faced by native languages in Latin America and the lack of language technologies faced by low-resource languages in Peru (Rogers and Campbell, 2015). This situation is mainly due to the lack of speech corpora, which are the raw material for the creation of language tools, are scarce and the few that exist are privately licensed; for this reason, they should be in the public domain to contribute to the development and revitalization of languages.

Around the world, there are some initiatives for the collection of corpora for low-resources languages and that are in the public domain that employ different methodologies and ways of collection. One of the most successful methodologies is crowdsourcing, i.e. native speakers volunteering to help in the construction of the resources. This methodology is supported by web tools or mobile applications, which can be massively used.

Our corpus collection tool is designed to expand organically to new native languages as community members record domain-specific base audios as prompts in the corpus collection. Unlike others (Common Voice (Ardila et al., 2019)), this tool does not use text to be read, as many native speakers of indigenous languages are illiterate in their native language. Therefore, our tool replaces reading texts with listening to audios. This subtle but important change facilitates corpus collection.

Native Languages

Peru is a multicultural country, mainly due to the presence of native first nations, these make up a total of 10% of the population. Because of this population there are still 48 native languages spoken, however, they are under the risk of extinction. These languages are facing some major issues like the lack of a unique grammar or writing system, lack of presence on the internet, lack of mass of expert linguists and lack of electronic resources. (Cardenas et al., 2018)

Quechua

Quechua (ISO 639-3 que) is a family of languages spoken in South America with about 10 million speakers, not only in the Andean regions but also in the valleys and plains connecting the Amazon jungle and the Pacific coast. Quechua languages are considered highly agglutinative with a subject-object-verb (SOV) sentence structure as well as mostly postpositional. Even though the classification of Quechua languages remains open to research (Heggarty et al., 2005; Landerman, 1992), recent work in language technology for Quechua (Rios, 2015; Rios and Mamani, 2014) have adopted the categorization system described by Torero (Torero, 1964). This categorization divides the Quechua languages into two main branches, QI (Glottolog quec1386) and QII (quec1388). Branch QI corresponds to the dialects spoken in central Peru, which are treated as one collective in this paper. QII is further divided in three branches, QIIA, QIIB and QIIC. QIIA groups the dialects spoken in Northern Peru, while QIIB the ones in Ecuador and Colombia. In this paper we work with QI (Central Quechua, Glottolog quec1386) and QIIC (Southern Quechua, Glottolog quec1389).

Aymara

he Aymara language (ISO 639-3 aym) belongs to the Aru linguistic family, is spoken by the Aymara people and although it is in a vital state (MINEDU, 2018), it is considered an endangered language (Adelaar, 2014). Aymara is spoken in four countries: Argentina, Bolivia, Chile and Peru. In Peru, it is the second most spoken native language after Quechua, according to the 2017 census conducted by the National Institute of Statistics and Informatics (INEI, 2017). It is an agglutinative language. Aymara has 3 short phonemic vowels /a/, /i/ and /u/, and 3 long ̈a /aa/, ̈ı /ii/ and ̈u /uu/. Also, it features 26 consonant phonemes, most of them aspirated occlusives ph [ph], th [th] and kh [kh]. In addition, the aspirated postalveolar affricate is signaled by the triplet chh [tSh] and an apostrophe is used to signal the occlusive and affricate ejective p’ [p’], t’ [t’], ch’ [ch’] and k’ [k’]. Like Spanish and Quechua it features the phonemes /ch, /ñ/ ̃n, and /L/ ll (MINEDU, 2021).

Shipibo Konibo

The Shipibo-Konibo people are one of the most influential communities in the Peruvian Amazon. They call themselves ”Jonikon”, which means ”real people”; they also adopted the exonym ”shipibo”. Their own language or ’joikon’, ’true language’ is now known as Shipibo-Konibo. This language belongs to the Panoan linguistic family, which is an important subject of study for many linguistic researchers in Peru (Adelaar, 2014; Zariquiey and others, 2006). Shipibo-Konibo is an agglomerative language, with a high use of common suffixes (130) plus some prefixes (13) for its word-formation process. Furthermore, the basic sentence order is SOV (subject-object-verb) as opposed to Spanish (SVO) (Valenzuela, 2003). This language is spoken by around 22 thousand people in 150 communities and is taught in almost 300 public schools (Sull ́on Acosta et al., 2013). The phonological repertoire of Shipibo consists of 16 consonants and 4 vowels. The vowels in Shipibo are characterized by the presence of two heights (high and low), among which it is important to point out the high central vowel, not rounded 1. In the consonant phonemes we find four labial [p], [b], and [m], nine coronal [t], [s], [ts], [n], [S], [y], [s ̧] and [r], two dorsal [k] and [w] and one global [h] (Martinez, 2009).

Corpus Creation

Like Common Voice, we used the crowdsourcing method, which is based on the massive help of volunteers for audio recordings. This methodology allowed us to collect as many audios as possible in a short time and with a small budget. We used two corpus collection applications (Huqariq, Tarpuriq) designed exclusively to record and validate respectively. Unlike the Common Voice platform, the volunteers do not have to read a sentence but listen to it. This last functionality is important for native languages of Peru, due to a large part of the native speaker population are illiterate.

The official dictionaries of each language described in this research were used. We used the official dictionaries issued by the Peruvian Ministry of Education, because the texts in the dictionaries are correctly written according to the official standard of each language. 4 native speaker linguists corrected, normalized and standardized the texts according to the grammar issued by the Ministry of Education and Ministry of Culture for each language. On the other hand, for the Southern Quechua sentences, a morphological analyzer (Rios, 2015) was used, which automatically standardizes according to the rules of the Ministries of Education and Culture.

Linguists who are native speakers of each language recorded their voices reading each of the selected texts. The recordings were made using the Tarpuriq application for Android, which has an audio recording module very similar to Huqariq application for Android (Camacho and Zevallos, 2020). The recordings were made in a controlled environment, mainly free of noise and interference of any kind. All recordings made by the linguists were stored in a folder called ”prompts” and folders named after their respective languages. All the recordings (prompts) made by the linguists are then entered into the Huqariq application so that they can be listened to by the volunteers to record their voices. Finally, the prompts were saved as 16-bit, single channel WAV audio files with a sampling frequency of 16 kHz.

For the collection of recordings (audio files) from native speakers (users), Huqariq was used. This application allowed native speakers to record their voices repeating the sentences they hear in the prompts mentioned above. The app assigns 200 sentences per user, this feature of Huqariq was developed in this research in order for users to have a goal and to be able to be rewarded when they achieve it. The recordings of the volunteers have the same technical information as the prompts. The recordings made by users were validated using 2 methods. The first method used an automated quality validation module that checks the noise, silence and duration of the recordings, this method was incorporated into the Huqariq application. The second method was performed by Tarpuriq, which allowed native linguists of the respective languages to validate the quality of the recordings through a voting system, this method is similar to the one used by Common Voice. Each recording must be voted 3 times, if a recording receives two positive votes, it will be marked as valid, on the contrary, if it receives two negative votes, it will be marked as invalid. Recordings marked as valid will be added to the final training, development and test corpus. This corpus is a July 2021 version of the corpus, which is the most updated, since due to the pandemic we have not been able to continue working on the validation of the corpus. Table 1 shows some statistics important .

Language	Number of Sentences	Volunteers	Total Hours	Validated Hours
Southern Quechua	800	480	340	180
Central Quechua	1171	20	20	20
Aymara	1900	8	15	14
Shipibo-Konibo	500	2	7	6

Table 1: Statistics of Huqariq corpus

Experiment

The following experiment demonstrates the potential of the Huqariq corpus for multilingual speech research for low-resource languages. For this experiment we used the corpus described in Table 1. We used the pre-trained model Wav2Vec2 (Baevski et al., 2020) which was trained with 600 hours of Spanish2. It is important to mention that we use a pre-trained model of Spanish because the languages in our corpus contain many borrowings from Spanish and this can improve the performance of the model. Moreover, we used the training setup from the public repository of which we obtained the pre-trained model. Table 2 shows the results of the Wav2Vec2 model for each trained language and the results of previous work.

Model	Southern Quechua	Central Quechua	Aymara	Shipibo-Konibo
Wav2letter++ +(DA)	31.48 22.75			
Wav2vec2 +CTC (subword)	28.73	41.15	59.81	72.15
+LM (decode)	23.19	36.37	52.6	67.47

Table 2: Performance results of the ASR models performed for each language using the CER metric.

Conclusion

We have presented Huqariq: a multilingual speech corpus of Peruvian native languages for the development of speech recognition tools. By using the crowdsourcing methodology and 2 mobile applications we have collected the largest speech corpus of native Peruvian languages. On the other hand, we have conducted some experiments on automatic multilingual speech recognition with the Huqariq corpus using the Wav2Vec2 model. This is the first time that speech recognition experiments have been performed for Central Quechua, Aymara and Shipibo-Konibo.