scientific data

OPEN

DATA DESCRIPTOR

A speech corpus of Quechua Collao for automatic dimensional emotion recognition

Rosa Y. G. Paccotacya-Yanque ✉, Candy A. Huanca-Anquise, Judith Escalante-Calcina, Wilber R. Ramos-Lovón & Álvaro E. Cuno-Parari

Automatic speech emotion recognition is an important research topic for human-computer interaction and affective computing. Over ten million people speak the Quechua language throughout South America, and one of the most known variants is the Quechua Collao one. However, this language can be considered a low resource for machine emotion recognition, creating a barrier for Quechua speakers who want to use this technology. Therefore, the contribution of this work is a 15 hours speech corpus in Quechua Collao, which is made publicly available to the research community. The corpus was created from a set of words and sentences explicitly collected for this task, divided into nine categorical emotions: happy, sad, bored, fear, sleepy, calm, excited, angry, and neutral. The annotation was performed on a 5-value discrete scale according to 3 dimensions: valence, arousal, and dominance. To demonstrate the usefulness of the corpus, we have performed speech emotion recognition using machine learning methods and neural networks.

Background & Summary

Studies have shown that AI is able to understand and express emotions, improving the quality and effectiveness of Human-Machine Interaction[1,2]. Speech emotion recognition (SER) is useful in affective computing applications such as patient in-home monitoring[3], early detection of psychiatric diseases and disorders[4–6], support diagnosis and treatment in military healthcare[7], recognize deceptive speech[8,9], or stress on it[10,11]. Furthermore, SER is used in call center conversations to categorize voice mail messages. Humans naturally can recognize emotions from facial expressions and speech. Experts claim that facial expressions are universal. However, speech is not; there are variations in the vocal signature for some emotions, e.g., anger. This situation is due to differences in the expression of emotions in different cultures and languages[12,13].

Quechua is spoken in seven South American countries (Peru, Ecuador, Colombia, Bolivia, Argentina, Chile, and Brazil). In Peru, it is considered a vital language because it was spoken before the diffusion of the Spanish language, and it is used across the country. However, many Quechua variants are in danger, given that there is a significant decrease in the importance of this language[14]. That is reflected in the lack of Quechua speech datasets annotated with emotions when compared, for example, to Spanish or English. Languages with data scarcity are known as low-resource languages (LRLs)[15]; in this sense, Quechua is a low-resource language.

To the best of our knowledge, no work in the literature focuses on creating a corpus in Quechua for SER. However, few proposals involve the creation of speech corpora for Quechua that deal mainly with automatic speech recognition (ASR). One of these works is Siminchik[16], a speech corpus for the preservation of Southern Quechua, intended to be used in speech recognition systems. It was created from radio programs and volunteers (native Quechua speakers) who made audio transcriptions. A Hidden Markov Model was used to perform ASR. Although Siminchik has a larger version[17] that was created by adding more data sources such as recordings of repetitions of audio prompts, readings of text prompts, and free speeches of native speakers, it still seems to be a work in progress (https://www.siminchikkunarayku.pe)[18]. Also, Chacca *et al.*[19] used Mel-Frequency Cepstral Coefficients (MFCC), Dynamic Time Warping (DTW), and K-Nearest Neighbor (KNN) in an ASR system to recognize numbers in Quechua. For this, isolated words were recorded from 30 native speakers, between men and women, counting from 1 to 10. However, this database was limited as it contained only 300 audios.

Universidad Nacional de San Agustín de Arequipa, School of Computer Science, Arequipa, Peru. ✉e-mail: rpaccotacya@unsa.edu.pe

	Gender	Age	Profession	Birthplace
Actress 1	F	43	Quechua Instructor	Puno
Actor 2	M	36	Quechua Instructor	Puno
Actress 3	F	49	Quechua Instructor	Cusco
Actor 4	M	28	Quechua Instructor	Cusco
Actress 5	F	45	Quechua Professor	Cusco
Actor 6	M	36	Quechua Instructor	Apurimac
Actress 7	F	24	Quechua Instructor	Puno

Table 1. Demographical information of actresses and actors.

Therefore, there is an emotion recognition gap for Quechua speakers. We aim to fill this gap by creating a speech corpus of Quechua Collao for automatic emotion recognition, evaluating its usefulness using machine learning techniques and deep neural networks, and making it publicly available. We selected the dimensional approach, based on valence, arousal, and dominance proposed by Mehrabian and Russell[20] to indicate people's state of feeling, as categorical emotions can be recovered from dimensional values[21]. Evaluation with machine learning techniques and neural networks will be helpful as a baseline for other researchers who want to study the recognition of speech emotions with supervised and unsupervised algorithms. Therefore, the creation of this corpus not only fills the emotion recognition gap for Quechua speakers but also opens the opportunity for new studies.

Methods

Script creation. Our script comprises 378 words and 1692 sentences, making a total of 2070 instances. These were extracted from different texts written in Quechua Collao, while most of them include a Spanish translation[22–36]. The complete script is divided into nine parts, each consisting of 42 words and 188 sentences. These parts correspond to 9 categorical emotions: happy, sad, bored, fear, sleepy, calm, excited, angry, and neutral. Most words and sentences were chosen according to the emotions that were used. For example, the sentence "Sumaqmi ñawiyki", which means "your eyes are pretty", was selected to represent happy emotion.

The script construction was done in five phases. In the first phase, all the documents from which all the words and sentences were extracted were investigated. The second phase was carried out by three people who built the script. These first two phases were carried out by people who speak Spanish natively but do not speak any variant of Quechua. In the third phase, an expert in the Quechua Collao language reviewed and corrected the entire script and provided other sources to replace some sentences per emotion. The fourth phase consisted of selecting these new sentences and replacing some old ones. Finally, the fifth phase was accomplished by a second Quechua language expert, who translated sentences from Spanish into Quechua Collao and made corrections to the entire script.

Recording sessions. The recording sessions were performed with semi-professional microphones and at semi noiseless spaces at the School of Computer Science - UNSA. The actors were mostly mid-age native Quechua speakers that were paid to record. The whole script (2070 instances) was planned to be recorded by each actor of a group of three women and three men in order to have balanced data in terms of gender. However, an actress (Actress 5) could not finish due to unexpected situations, so her work was completed by another actress (Actress 7). Table 1 shows information related to actors.

A session consists of one emotion's recording per actor (230 instances). Previous to the recording, the actors were given the scripts and asked to rehearse with the linguist expert. In each session, the actor/actress is asked to read the script while interpreting the target emotion, and its interpretation is evaluated. If the evaluator decides the performance or pronunciation is inadequate, then the recording of the instance is repeated, asking the actor/actress to remember situations in the past or giving them hypothetical situations so emotions can be elicited. To avoid performance fatigue, each session was divided into five groups of 46 instances, and after recording each group, a pause of around 5 minutes was done.

The audios were recorded using one of the following microphones: Emita USB Studio Microphone GXT 252 and Blue Snowball USB Microphone. The software used to record was Audacity, with a sample rate 44KHz and one recording channel (mono). The scene setup for recordings is shown in Fig. 1.

Segmentation. There were two segmentation stages. First, an automatic segmentation was performed using the software Praat. Each WAV file obtained from the recording session was divided into small segments using a script that splits the original audio file after detecting silence. However, this process is not flawless, and a manual segmentation had to be performed. Each audio segment was listened to by a person who verified if it was correct or not. If not, the segment was lengthened or shortened to avoid cutting off words and to delete noise or silence at the beginning or end of the segment. In addition, some fragments were merged when a word or sentence was split incorrectly. After the automatic and manual segmentation, the resulting audios were assigned a random name so they could be labeled without previous information on which emotion they belonged to by the annotators.

Annotation. Two men and two women were employed and paid to annotate the audio labels. They are Quechua Collao native speakers and Quechua instructors, ages ranging from 27 to 46.

Fig. 1 Scene setup for recordings.

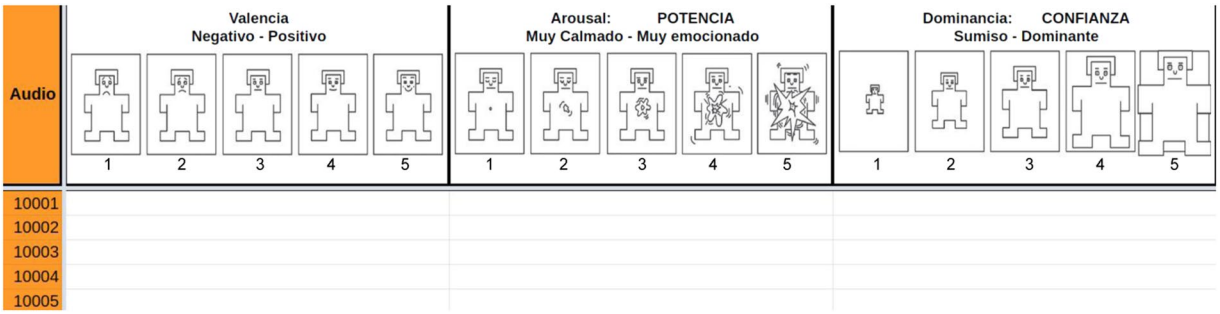

Fig. 2 Example of the sheet used by the annotators to label the audios.

Each annotator labeled the 12420 audios over a 4-week period by assigning valence, arousal, and dominance values. Each week, 3105 audios were released to be annotated. It must be noted that we only considered five days a week dedicated to annotation. A methodology was recommended to ensure the annotation quality, suggesting about 6 hours a day to annotate, preferably in two 3-hour groups, while taking a 10-minute pause after 1.5 hours. During the last week, one of the female annotators had to be replaced by another female Quechua instructor due to some external inconveniences.

The annotation process was performed using a sheet for each annotator, where they had to write the valence, arousal, and dominance values for each audio. A scale of 1 to 5 was used, and visual aid was provided using self-assessment manikins (SAMs)[37]. Figure 2 shows an example of the sheets used for annotation, where the first column shows the audio filename, and the last three columns correspond to valence, arousal, and dominance. In the header, 5 SAMs are shown for each emotional dimension, with a brief description in Spanish: Negative - Positive for valence, Very calm - Very excited for arousal, and Submissive - Dominant for dominance.

Ethics declaration. All procedures performed in this study were approved by the KUSISQA project research committee in accordance with the ethical standards of the National University of San Agustin. Likewise, individual written informed consent was obtained from all participants involved in the study, where the actors and actresses were informed that their voices would be freely shared anonymously.

Data Records

The corpus is publicly available at Figshare[38]. It contains 15 h 15 min of audio divided into 12420 segments of audio. Table 2 summarizes its content.

The corpus[38] can be downloaded as a zip file that contains 4 directories. All audio segments in WAV format are found in the **Audios** folder; each audio is randomly named by numbers ranging from 10001 to 22420. This action was carried out to avoid bias. Detailed information of each audio is found in a file inside the **Data** folder, in a file named *Data.csv*. This file is made up of five columns, where:

- The *Audio* column contains the name of the WAV files.
- The *Emotion* column represents the categorical emotion of the audio segment.
- The *Actor* column contains the code of the actor who performed and interpreted the audio segment. The code comprises an *a* and a number from 1 to 6 representing the six actors (example: a2).
- The *File* column contains the original name of each audio segment, this name is divided into two segments by an underscore: The first represents the code of the actor, and the second segment is composed of a letter and a number; the letter represents the emotion (H = Happy, T = Sad, B = Bored, F = Fear, S = Sleepy, C = Calm, E = Excited, A = Angry, and N = Neutral) and the number represents the position of the audio segment. For

Corpus features	
Number of audio segments	12420
Total segments length	15 h 15 m 15 s
Dimensions	Valence, Arousal, and Dominance
Annotations scale	1, 2, 3, 4, 5
Number of speakers	7
Gender of speakers	4 women and 3 men
Age range of speakers	24–49 years

Table 2. Corpus[38] summary.

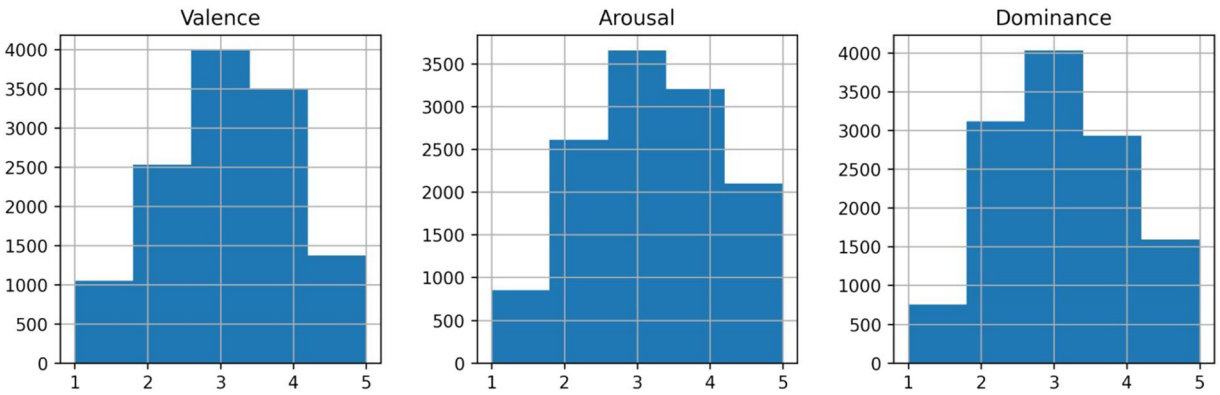

Fig. 3 Histograms for each dimensional attribute in the Quechua Collao corpus.

example, the name a2_B159.wav indicates that the segment belongs to actor 2 (a2), that it belongs to the emotion bored (B), and that it is the 159th segment in this set.
- *Duration (s)* column indicates the duration in seconds of each audio segment.

The annotations for each audio segment are found in the **Labels** folder, and the emotional dimensions of valence are found in the *Valence.csv* file, which is made up of 5 columns: the first contains the name of the audio segment, and the other four store the label made by each annotator. For example, the second column contains the labels made by annotator 1, represented by the code N1. The values for the emotional dimension of arousal and the emotional dimension of dominance are found in the files *Arousal.csv* and *Dominance.csv*, respectively, and have the same structure as the *Valence.csv* file.

The general average of each emotional dimension is found in the *Labels.csv* file, also within the **Labels** folder, where the first column represents the name of the audio segment, the second, third, and fourth columns contain the average of the labels made by the four annotators of the dimensions emotional valence, arousal, and dominance respectively.

Finally, the **Script** folder contains *Script.xlsx*, which is the script used for the corpus[38] creation. This file comprises 9 sheets for the 9 categorical emotions used. Each word and sentence has an ID that employs the same notation as that of the *File* column of *Data.xlsx*, explained previously, without considering the actor prefix. For example, T159 is an ID that corresponds to the 159th sentence of the Sad emotion.

Technical Validation

Below, we present three validations of our corpus[38]: Emotional Diversity, Annotation Consensus, and Machine Validation.

Emotional diversity. To evaluate the emotional diversity, we created histograms for each dimension that depict instances' distribution in labels from 1 to 5 (Fig. 3). As can be seen, the corpus[38] has an unbalanced emotional content (valence - arousal), a common problem in most data sets for SER[39]. Data imbalance is also observed for dominance. This problem could be solved by adding more annotators.

Annotation consensus. The reliability of the labels between annotators was calculated using Cronbach alpha coefficients. Table 3 shows the results; as can be seen, there is a higher agreement for arousal and dominance than for valence. It must be noted that all 12420 audios were used to calculate the Cronbach alpha coefficients.

Machine validation. To validate the usefulness of this Quechua Collao corpus[38], experiments for dimensional emotion recognition were carried out using machine learning methods and neural networks. These experiments establish a baseline standard for future research.

Valence	Arousal	Dominance
0.513	0.619	0.625

Table 3. Cronbach's alpha coefficients.

LLDs	Intensity, alpha ratio, hammarberg index, spectral slope 0–500 Hz, spectral slope 500–1500 Hz, spectral flux, 4 MFCCs, f_o, jitter, shimmer, Harmonics-to-Noise Ratio (HNR), Harmonic difference H1-H2, Harmonic differ- ence H1-A3, F1, F1 bandwidth, F1 amplitude, F2, F2 amplitude, F3, and F3 amplitude.
Feature vector	Mean (of LLDs), std (of LLDs).

Table 4. GeMAPS low-level descriptors (LLDs) and feature vector structure. Adapted from Atmaja and Akagi's work[41].

The machine learning (ML) methods used are Support Vector Regression (SVR), K-neighbors Regression (KNR), and Random Forest Regression (RFR). They were implemented using Scikit-Learn[40].

The neural network models used are Multilayer Perceptron (MLP), Long Short-Term Memory (LSTM) network, and Convolutional Neural Network (CNN), whose implementation details can be found in Atmaja and Akagi's work[41]. The hyperparameters were adjusted to the Quechua Collao corpus[38] for each model.

All the parameters for these methods were obtained after extensive testing to optimize the mean of the Concordance Correlation Coefficient (CCC)[42] scores for valence, arousal, and dominance.

Recall that the final label for each audio track was calculated as the average of the annotators' individual labels.

A feature vector was obtained from each audio by using the Geneva Minimalistic Acoustic Parameter Set (GeMAPS)[43] and the procedure done by Atmaja and Akagi[41]. Twenty-three low-level descriptors (LLDs), presented in Table 4, were computed on a frame-based level. Mean and standard deviation were calculated over the feature set to attain a vector with 46 elements for each audio.

After obtaining the feature vectors for all audios, they were separated into three sets: 60% for training, 20% for validation, and 20% for testing. All methods use these sets. Their results are measured using the CCC criteria.

K-Neighbors regressor. KNR is the regression based on k-nearest neighbors (KNN). Since KNN has been used in several SER categorical challenges[44–46]. In this case, we used KNR to validate a multidimensional problem since we need to intuitively approximate the association between variables and the continuous result. For the implementation we used the Sklearn library and the following parameters: *8* as the number of neighbors, *distance* as the weight function, *Minkowski* as the distance metric and *1* as the power parameter.

Support vector regressor. SVR is a widely used method in approaches to tackle SER[47–49], as well as in SER challenges, like the AudioVisual Emotion Challenge (AVEC), to establish a baseline[50].

We used the Nu Support Vector Regression (NuSVR) implementation with the following parameters: *nu* was set to 1, *C* was 172600, the kernel was rbf and *gamma* was *scale*.

Random forest regressor. A Random Forest is an estimator that works by averaging the prediction of different decision trees. RFR has been used in different works of SER[51–53], and also as a baseline for a new audio-visual dataset[54]. For the implementation, we used Sklearn with 100 trees.

Multilayer perceptron. The MLP model was trained with five layers of 256,128, 64, 32, and 16 nodes per layer and ReLU activations. The optimizer was Adam, with a learning rate of 0.001 and a batch size of 32.

Long short-term memory network. The LSTM model used an Adam optimizer, a batch size of 32 and 3 layers with 512, 256, and 128 nodes.

Convolutional neural network. The CNN used for recognizing speech emotions is made up of 5 layers. The first layer uses 256 neurons, and these are reduced by half in each of the subsequent layers (256, 128, 64, 32, 16). The activation function ReLU was used in the five layers, and the batch size used was 32. An Adam optimizer is used to adjust the learning rate during the training process.

The MLP, CNN, and LSTM models were trained for a maximum of 180 epochs with an early stopping with patience of 10 epochs monitored on the validation loss.

The CCC scores for the ML methods are presented in Table 5. RFR has the best mean on the CCC scores for each dimension, and arousal and dominance have higher CCC scores than valence.

Table 6 shows the CCC scores of the neural network methods. The LSTM model performs better than CNN and MLP in average, while the tendency observed in the ML methods also occurs, as valence has a lower CCC score when compared to arousal and dominance.

Method	Validation set				Test set			
	Valence	Arousal	Dominance	Mean	Valence	Arousal	Dominance	Mean
SVR	0.401	0.646	0.698	0.582	0.427	0.625	0.684	0.578
KNR	0.324	0.479	0.520	0.441	0.294	0.508	0.538	0.447
RFR	**0.485**	**0.699**	**0.735**	**0.639**	**0.492**	**0.684**	**0.717**	**0.631**

Table 5. Results of CCC scores for ML methods.

Method	Validation set				Test set			
	Valence	Arousal	Dominance	Mean	Valence	Arousal	Dominance	Mean
MLP	0.612	0.748	0.769	0.710	0.636	0.756	0.775	0.723
LSTM	**0.627**	0.749	**0.778**	**0.718**	**0.648**	**0.764**	0.782	**0.731**
CNN	0.612	**0.751**	0.766	0.710	0.632	0.747	**0.785**	0.721

Table 6. Results of CCC scores for neural network methods. These scores were averaged over 10 runs of each algorithm.

In general, neural network methods have better results than ML methods, while CCC scores for valence consistently are lower than those for arousal and dominance. This fact reflects the Cronbach alpha coefficients shown in Table 3, where the valence labels have the lowest agreement among the annotators.

Usage Notes

The individual labels are also provided along with the averaged ones, making it possible to apply a different method to calculate the final labels for each audio.

The corpus[38] is built over the variant Collao of Quechua. Thus, this variant has noticeable differences from other variants and should not be used to generalize Quechua.

The main limitation is the emotional imbalance of the corpus[38], which can lead to low performance in SER algorithms for labels of instances with low frequency. Furthermore, we must note that the recordings are performed in a controlled environment with only six mid-age speakers acting established emotions. Therefore, given these limitations, the recordings provide prototypical insights for studying emotions but they cannot fully represent the emotional expressions of all Quechua speakers that are observed in real life.

Code availability

Code and data splits for baseline algorithms are available at Github, in https://github.com/qccData/qccCorpus.

Received: 1 August 2022; Accepted: 21 November 2022;
Published online: 24 December 2022

References

1. Becker, C., Kopp, S. & Wachsmuth, I. *Why Emotions should be Integrated into Conversational Agents*, chap. **3**, 49–67 (John Wiley & Sons, Ltd, Hoboken, NJ, USA, 2007).
2. Ball, G. & Breese, J. *Emotion and Personality in a Conversational Agent. 189–219.* (MIT Press, Cambridge, MA, USA, 2001).
3. Mano, L. Y. *et al.* Exploiting iot technologies for enhancing health smart homes through patient identification and emotion recognition. *Computer Communications* **89-90**, 178–190, https://doi.org/10.1016/j.comcom.2016.03.010 (2016).
4. Tacconi, D. *et al.* Activity and emotion recognition to support early diagnosis of psychiatric diseases. *2nd International Conference on Pervasive Computing Technologies for Healthcare 2008, PervasiveHealth* 100–102, https://doi.org/10.1109/PCTHEALTH.2008.4571041 (2008).
5. Low, L.-S. A., Maddage, N. C., Lech, M., Sheeber, L. B. & Allen, N. B. Detection of clinical depression in adolescents' speech during family interactions. *IEEE Transactions on Biomedical Engineering* **58**, 574–586, https://doi.org/10.1109/TBME.2010.2091640 (2011).
6. Stasak, B., Epps, J., Cummins, N. & Goecke, R. An investigation of emotional speech in depression classification. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 485–489, https://doi.org/10.21437/Interspeech.2016-867 (2016).
7. Tokuno, S. *et al.* Usage of emotion recognition in military health care. In *Defense Science Research Conference and Expo (DSR)*, 1–5, https://doi.org/10.1109/DSR.2011.6026823 (2011).
8. Amiriparian, S., Pohjalainen, J., Marchi, E., Pugachevskiy, S. & Schuller, B. W. Is deception emotional? an emotion-driven predictive approach. In *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2011–2015, https://doi.org/10.21437/Interspeech.2016-565 (2016).
9. Zuckerman, M., DePaulo, B. M. & Rosenthal, R. Verbal and nonverbal communication of deception. *Advances in Experimental Social Psychology* **14**, 1–59, https://doi.org/10.1016/S0065-2601(08)60369-X (1981).
10. Hansen, J. H. L. Analysis and compensation of speech under stress and noise for environmental robustness in speech recognition. *Speech Communication* **20**, 151–173, https://doi.org/10.1016/S0167-6393(96)00050-7 (1996).
11. Zhou, G., Hansen, J. H. L. & Kaiser, J. F. Nonlinear feature based classification of speech under stress. *IEEE Transactions on Speech and Audio Processing* **9**, 201–216, https://doi.org/10.1109/89.905995 (2001).
12. Swain, M., Routray, A. & Kabisatpathy, P. Databases, features and classifiers for speech emotion recognition: a review. *International Journal of Speech Technology* **21**, 93–120, https://doi.org/10.1007/s10772-018-9491-z (2018).
13. Byun, S.-W. & Lee, S.-P. A study on a speech emotion recognition system with effective acoustic features using deep learning algorithms. *Applied Sciences* **11**, https://doi.org/10.3390/app11041890 (2021).

14. Ministerio de Cultura de Perú. Quechua. *Base de Datos de Pueblos Indígenas u Originarios* https://bdpi.cultura.gob.pe/lenguas/quechua (2012).
15. Magueresse, A., Carles, V. & Heetderks, E. Low-resource languages: A review of past work and future challenges. Preprint at https://arxiv.org/abs/2006.07264 (2020).
16. Camacho, L., Zevallos, R., Cardenas, R. & Baquerizo, R. Siminchik: A speech corpus for preservation of southern quechua. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC'18)*, https://doi.org/10.5281/zenodo.3595354 (2019).
17. Camacho, L., Vergara, M. & Zevallos, R. On the building of the large scale corpus of southern qichwa. In *Latin American and Iberian Languages Open Corpora Forum* (2017).
18. Guzman, Y., Tavara, A., Zevallos, R. & Vega, H. Implementation of a bilingual participative argumentation web platform for collection of spanish text and quechua speech. In *2021 International Conference on Electrical, Communication, and Computer Engineering (ICECCE)*, 1–6, https://doi.org/10.1109/ICECCE52056.2021.9514251 (2021).
19. Chacca, H., Montufar, R. & Gonzales, J. Isolated automatic speech recognition of quechua numbers using MFCC, DTW and KNN. *International Journal of Advanced Computer Science and Applications (IJACSA)* **9**, https://doi.org/10.14569/IJACSA.2018.091003 (2018).
20. Mehrabian, A. & Russell, J. A. *An approach to environmental psychology* (The Massachusetts Institute of Technology, USA, 1974).
21. Russell, J. A. Affective space is bipolar. *Journal of personality and social psychology* **37**, 345–356, https://doi.org/10.1037/0022-3514.37.3.345 (1979).
22. Cusihuamán, G. A. *Gramática Quechua*, vol. 1 (Centro Bartolomé de las Casas, Cusco, Perú, 2001).
23. Herrera, A. *et al*. *Guías pedagógicas del sector lengua indígena Quechua material de apoyo para la enseñanza* (Ministerio de Educación de Chile, Santiago de Chile, Chile, 2012).
24. Chuquimamani-Valer, N. R., Alosilla-Morales, C. G. & Choque-Valer, V. *Qullaw qichwapa simi qullqan: A-Y* (Ministerio de Educación, Lima, Perú, 2014).
25. Aranda-Escalante, M. *Manual para el empleo del Quechua Cusco Collao en la administración de justicia* (Ministerio de Cultura, Lima, Perú, 2015).
26. Hancco-Mamani, N. A. *et al. Runa simi qillqay yachana mayt'u* (Ministerio de Cultura, Lima, Perú, 2013).
27. Pinto-Tapia, M. *Didáctica Quechua I - DRE Apurímac EBI-Rural* (Direcciòn Regional de Educación de Apurímac, Apurímac, Perú, 2005).
28. Kindberg, E. & de Kindberg, K. L. *Palabras útiles en el quechua de Caylloma* (Instituto Lingüístico de Verano, Apurímac, Perú, 1985).
29. Sullca-Peña, A. *Kuska yachasunchik. Cuaderno de trabajo y folder - inicial 4 años Quechua Collao*, 4 edn (Ministerio de Educación, Lima, Perú, 2020).
30. Chuquimamani-Valer, N. R. *Yachakuqkunapa Simi Qullqa - Qusqu Qullaw Qhichwa Simipi* (Ministerio de Educación, Lima, Perú, 2005).
31. Llamoja-Tapia, M. *Simikunapi, Kawsaykunapi, Sapsikunapi Tukuy niraq Yachachina Umalliq Iskay Simipi Kawsaypura Yachachiy Umalliq* (Gráfica Biblos S.A., Lima, Perú, 2021).
32. Ortiz-Vásquez, R. *Kunan punchaw runasimita istudyasaq* (Universidad Nacional Mayor de San Marcos, Lima, Perú, 2017).
33. Rodríguez, A. *Quechua de Cusco - Collao* (Lima, Perú, 2021).
34. Zariquiey, R. & Córdova, G. *Qayna, kunan, paqarin. Una introducción práctica al quechua chanca* (Pontificia Universidad Católica del Perú, Lima, Perú, 2008).
35. Cahuana, R. *Manual de gramática Quechua Cusco-Collao* (Cusco, Perú, 2007).
36. Instituto Nacional de Cultura. *Huchuy Pumacha - Pumita* (Instituto Nacional de Cultura, Cusco, Perú, 2003).
37. Fischer, L., Brauns, D. & Belschak, F. *Zur Messung von Emotionen in der angewandten Forschung* (Pabst Science Publishers, Lengerich, Alemania, 2002).
38. Paccotacya-Yanque, R. Y. G., Huanca-Anquise, C. A., Escalante-Calcina, J., Ramos-Lovón, W. R. & Cuno-Parari, A. E. Quechua collao corpus for speech emotion recognition. *Figshare* https://doi.org/10.6084/m9.figshare.20292516 (2022).
39. Lotfian, R. & Busso, C. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing* **10**, 471–483, https://doi.org/10.1109/TAFFC.2017.2736999 (2019).
40. Pedregosa, F. *et al*. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
41. Atmaja, B. T. & Akagi, M. Deep multilayer perceptrons for dimensional speech emotion recognition. In *2020 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 325–331, https://doi.org/10.48550/ARXIV.2004.02355 (2020).
42. Lin, L. I.-K. A concordance correlation coefficient to evaluate reproducibility. *Biometrics* **45**, 255–268, https://doi.org/10.2307/2532051 (1989).
43. Eyben, F. *et al*. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing* **7**, 190–202, https://doi.org/10.1109/TAFFC.2015.2457417 (2016).
44. Umamaheswari, J. & Akila, A. An enhanced human speech emotion recognition using hybrid of prnn and knn. In *2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, 177–183, https://doi.org/10.1109/COMITCon.2019.8862221 (2019).
45. Al Dujaili, M. J., Ebrahimi-Moghadam, A. & Fatlawi, A. Speech emotion recognition based on svm and knn classifications fusion. *International Journal of Electrical and Computer Engineering* **11**, 1259–1264, https://doi.org/10.11591/ijece.v11i2.pp1259-1264 (2021).
46. Aljuhani, R. H., Alshutayri, A. & Alahdal, S. Arabic speech emotion recognition from saudi dialect corpus. *IEEE Access* **9**, 127081–127085, https://doi.org/10.1109/ACCESS.2021.3110992 (2021).
47. Atmaja, B. T. & Akagi, M. Multitask learning and multistage fusion for dimensional audiovisual emotion recognition. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4482–4486, https://doi.org/10.1109/icassp40776.2020.9052916 (IEEE, 2020).
48. Han, J., Zhang, Z., Ringeval, F. & Schuller, B. W. Prediction-based learning for continuous emotion recognition in speech. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5005–5009, https://doi.org/10.1109/ICASSP.2017.7953109 (2017).
49. Ortega, J. D. S., Cardinal, P. & Koerich, A. L. Emotion recognition using fusion of audio and video features. In *2019 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, 3847–3852, https://doi.org/10.1109/SMC.2019.8914655 (IEEE, 2019).
50. Schuller, B. W. Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Commun. ACM* **61**, 90–99, https://doi.org/10.1145/3129340 (2018).
51. Zvarevashe, K. & Olugbara, O. O. Recognition of cross-language acoustic emotional valence using stacked ensemble learning. *Algorithms* **13**, https://doi.org/10.3390/a13100246 (2020).
52. Maithri, M. *et al*. Automated emotion recognition: Current trends and future perspectives. *Computer Methods and Programs in Biomedicine* **215**, 106646, https://doi.org/10.1016/j.cmpb.2022.106646 (2022).
53. Deshpande, G., Viraraghavan, V. S., Duggirala, M. & Patel, S. Detecting emotional valence using time-domain analysis of speech signals. In *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, 3605–3608, https://doi.org/10.1109/EMBC.2019.8857691 (2019).
54. Kossaifi, J. *et al*. Sewa db: A rich database for audio-visual emotion and sentiment research in the wild. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **43**, 1022–1040, https://doi.org/10.1109/TPAMI.2019.2944808 (2021).

Acknowledgements
This research was carried out as part of the Kusisqa project, whose funder is Proyecto Concytec - Banco Mundial, Mejoramiento y Ampliación de los Servicios del Sistema Nacional de Ciencia, Tecnología e Innovación Tecnológica with grant reference number Contract N° 014-2019-FONDECYT-BM-INC.INV. The authors acknowledge the contribution of the School of Computer Science, UNSA, for allowing us to use its facilities for recording sessions and for granting financial support. We are also immensely grateful to the actors, actresses, annotators and Quechua language experts.

Author contributions
W.R.L. and A.C.P. conceived the idea, supervised the activities, revised the manuscript and, as the board of the Kusisqa project, provided guidelines for study procedures. J.E.C., R.Y.G.P.Y. and C.A.H.A. designed the script, designed, prepared, and conducted the data collection, pre-processed and constructed the corpus[38], designed, and conducted the technical validation, and wrote the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Correspondence and requests for materials should be addressed to R.Y.G.P.-Y.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.