WBG: Words by Gestures
Emergency Line for hearing impaired community

André Torneiro
2Ai – School of Technology, IPCA, Barcelos, Portugal
a17636@alunos.ipca.pt

Abstract

Receiving emergency calls from people having a hearing disability is a major challenge because they use signs to communicate. To solve this problem, we propose a novel approach capable to translate Portuguese Sign Language (LGP) to Portuguese Language (LP) and LP to LGP. To translate LGP to LP an all-body 3D Keypoint Detector was used to extract joints coordinates from the video sequence which will serve as input to the proposed model. The proposed neural network is a Bi-LSTM that gave as output a Portuguese textual representation of the gesture. Then we convert text to speech using gTTS API. To translate Portuguese speech to LGP we propose a CNN model to exploit acoustic information from speech and Google WEB Speech API is used to convert speech to text. Finally, a skeleton is generated to perform the representative gesture from input audio. To train and evaluate the proposed models two innovative datasets were created. One dataset contains videos representing the gestures and was used in Bi-LSTM model, the other dataset contains audio files and was used in CNN model. During the train, CNN model achieved a mean test accuracy of 87% The Bi-LSTM model achieved a mean test accuracy of 99%. Both models have been tested in real-time and have a very good performance.

Keywords—*Deep Learning, Sign Language, Speech to text, Text to speech, 3D Keypoints, Sound analysis.*

1. Introduction

Currently, more than 1.5 billion people across the globe have some degree of hearing loss [1], and in 2050 World Health Organization (WHO) estimates that it can reach 2.5 billion people, so we need to be prepared to communicate with the deaf community. The language that deaf people use to communicate with each other is Sign Language however, this type of language isn't universal and changes from country to country.

Based on the data from Inquérito Nacional de Saúde com Exame Físico (INSEF) [2]**,** this problem affects 23.7% of all population which demonstrates that is a big problem in Portugal and raises a question.

"If they can't hear, how can we communicate with them?"

The answer to this question is easy because there is the Portuguese Sign Language, which is one of three official languages in Portugal since 1997 [3]. This language is characterized by gestures performed through facial expressions, body movement, mouth, and tongue which represent words. The gestures used in sign language can be static or in movement, normally the alphabet is static, and words are performed by movement. Like sign language across the globe changes from country to country, also in the same country there are different gestures to represent the same word. An example of this can be seen in **Figure 1**.

One of the main problems of the deaf community reveals itself when they need help from emergency lines, e.g. **112**, because they can't communicate with the other person and vice versa. To solve this problem in 2016 an App was promised by the Portuguese Government, however, this project took two years to be ready because there weren't sufficient sign language speakers to translate what was being said [4]. Finally, in 2019 **MAI112** [5] was available, with seven LGP speakers and allows Video conference with simultaneous LGP translation, Geolocation, and text messages. Due to the pandemic situation, recently also Serviço Nacional de Saúde (SNS) launched a similar system [6], but they have only six LGP speakers.

When we compare the number of LGP speakers available in [5], [6] with the number of people with hearing impairment, we note that are few for so many people.

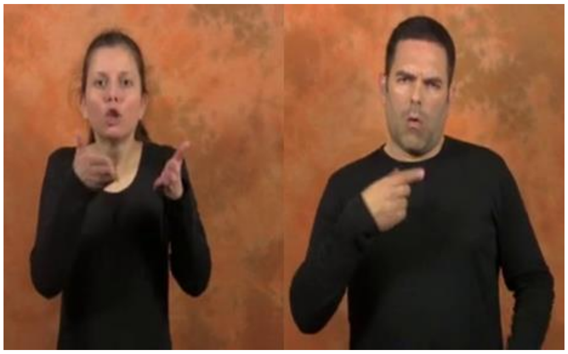

Figure 1: Example of two different initial movement in the same sentence "Posso ajudar-te?" (Retired from [7]).

And even if we think about the probability of more than six or seven people call to these emergency lines, at the same time, it is very high. So, to solve this problem we propose a system called WBG which is a new approach of artificial intelligence capable of integrating deaf community on emergency lines without a physical interpreter to make the communication. The main contributions of this work are the following:

- We propose a novel approach to Sign Language recognition in video.
- We are capable to distinguish the difference in the same sentence with different punctuation.
- We create two new datasets, one to classify intonation and the other to identify gestures.
- We provide a solution without any cost, when compared with others already existent.
- All system can perform in real-time.

The remainder of this paper is organized as follows. Section 2 presents related work. Section 3 presents all methodologies used to create this system. Section 4 and 5 reports the experimental settings and discusses the developments, advantages, and limitations in this system.

2. Related Work

In this section, we present a State-of-Art that summarizes the previous work done in Portugal related with Sign Language, then we talk about Sign Language Recognition with artificial intelligence approaches, and the last topic is related with speech analysis.

2.1. Portuguese Sign Language

Nowadays, some platforms are being created to help the deaf community, yet these are very similar to [5], [6], which raises the same problem for all, lack of LGP speakers. The most recent project is SERVIIN [8] which aims to enable a communication bridge with the deaf community and many companies like: GALP, EDP, APAV, and many others. This system receives video call from a deaf person with a cost of 1 cent/ minute, if he calls from a smartphone, or free if the call is made directly on his website. After the interpreter answers the call, he contacts the company desired and translates what is being said by the deaf person to the company (LGP to LP) and the reverse (LP to LGP).

Not only companies are interested in having the capability to integrate the community, but also the school community wants to have this capability. One of the reasons for schools to be interested in those types of systems is connected to the fact that they want to integrate all the children. In this direction, in 2015, Marcelo Norberto *et.al.* proposed a work called Virtual Sign [9] which aims to provide a virtual sign to assist the communication using bi-directional translation with deaf students in the classroom. This project is a game based on an avatar for making gestures and has three levels to a deaf person to learn to communicate, such as: how to make alphabet signs (level 1), represent words (level 2), and build a complete sentence (level 3).

Recently, important websites [10], [11], and metro station [12] add a Sign Language Avatar to their installations. The avatar used in [12] is based on the project from [9] but with the capability to speak six different languages. Also, in 2020, Paulo Fernandes [13] created an avatar to translate the Portuguese Language to Portuguese Sign Language, however, all avatars used by [10], [11], [12], [13] have a limitation since they can’t understand the context of the sentence, and perform signs word by word and not in a complete sentence. Since in LGP, the same word has different gestures dependent on the context (to see an example consult [14]) or a sentence has only one movement, so many times avatars make the wrong gestures.

With the development of artificial intelligence, new approaches were proposed, the simplest ones are [15], [16], [17], [18], and all are based on Kinect Sensor to extract keypoints of pose and Leap Motion to identify spatial coordinates from hands. Relatively to approaches with Machine Learning or Deep Learning, in 2019, S. Ferreira [19] implemented a system capable to recognize alphabet signs with a glove that has sensors to read hand position. This approach was tested in a neural network getting an accuracy of 78.27% and 71.16% when used by an LGP speaker and a non-LGP speaker, respectively. The best results were obtained when she made the same tests using a K-nearest neighbors algorithm reaching 93.31% and 89.66% when used by a LGP speaker and a non-LGP speaker, respectively.

Pedro M. Ferreira *et.al.*[20] are the authors of one of the major works done and that we consider one of the main references to this work, in Portugal. They started their research in 2014 by creating a novel dataset [21] that contains two topics: signLangDB (1), corpLangDB (2). Both topics were recorded in video format, but there is a difference between (1) and (2). The difference is that (1) contains 182 isolated signs, performed by a total of 15 LGP speakers, representing the alphabet and the numbers as well as nouns, pronouns, verbs, or common expressions, some performed with one hand and others with both, and 40 common sentences. The (2) contains a duo-interaction between deaf and/or hearing people, recorded with 13 volunteers, and enable the possibility of performing studies that analyze dialogue relationships. In 2019, they proposed DeSIRe [22] to learn a distribution about latent representations, independent of the signer’s identity. This solution only makes sense if we look at the full image and not just for the keypoints as we propose.

2.2. Sign Language with artificial intelligence

To make an easy and mutual communication between the hearing-impaired and the hearing communities, building a robust system capable of translating the spoken languages to sign languages and vice versa is fundamental. To this end, sign language recognition and production are two necessary parts for making such a two-way system.

2.2.1. Sign Language Recognition (SLR)

Proposed systems in sign language recognition generally map signs into the spoken language in the form of text transcription, however, that type of systems has critical challenges [23]. One of them is the visual variability of signs, which is affected by handshape, palm orientation, movement, location, facial expressions, and other non-hand signals. These differences in sign appearance produce a large intra-class variability and low inter-class variability. This makes it hard to provide a robust and universal system capable of recognizing different sign types.

Older approaches like [24], [24]–[29] trained and evaluated their models on either private or small-scale datasets with less than one hundred words. These approaches mainly consist in feature extraction, temporal-dependency modeling, and classification. Previous works first employ different hand-crafted features to represent static hand poses, such as SIFT-based features [30]–[32], HOG-based features [33]–[35] and features in the frequency domain [36], [37]. Hidden Markov Models (HMM) are then employed to model the temporal relationships in video sequences. Dynamic Time Warping (DTW) [38] is also exploited to handle differences of sequence lengths and frame rates. Classification algorithms, such as Support Vector Machine (SVM) [39], are used to label the signs with the corresponding words.

Recent approaches like [40]–[49], instead of looking at the full image, they only focus on skeleton keypoints. These approaches can be compared with Human Activity Recognition (HAR) like [50]–[53], yet, these methods only extract pose keypoints, and for SLR this isn't enough because hands and facial expression are the most important features, being able to consider pose almost unnecessary.

Normally, the approaches that use keypoints extraction are based on models from [54], [55]. These models first extract body bounding boxes, and then extract the keypoints. Recently, other detectors are proposed, such as, [56], [57], [58], [59].

2.2.2. Sign Language Production (SLP)

In contrast to SLR, SLP is still a very challenging problem involving an interpretation between visual and linguistic information [60]. The main difficulty of these systems is to generate the corresponding sign digit, word, or sentence from a text or voice in spoken language. This difficulty is resultant from the challenge of corresponding grammatical rules and linguistic structures of the sign language. Recent works on this topic are divided into two categories: Avatar approaches and NMT approaches.

Avatar approaches are based on 3D animated models, which can be stored more efficiently compared to videos. This technique can be programmed to be used in different sign languages. A common method to transfer body movement into the avatar is [61]. As mentioned before, unnatural movements, and missing non-manuals information, such as eye movement and facial expressions, are some challenges of the avatar approaches. To solve these problems, recent works focus on annotating non-manual information such as face, body, and facial expression [62], [63].

NMT approaches (neural machine translators) are a practical methodology for translating from one language to another. Sign Language Translation (SLT) aims to generate spoken language translations from sign language considering different word orders and grammar. Stoll et al. [60] proposed a hybrid model to automatic SLP using NMT, GANs, and motion generation. The proposed model generates sign videos from spoken language sentences with a minimal level of data annotation for training. This model first translates spoken language sentences into sign pose sequences. Then, a generative model is used to generate sign language video sequences.

2.3. Speech Recognition

Speech recognition and verification have increased visibility and significance in society as speech technology, audio content, and e-commerce continue to expand. There is an ever-increasing need to search for audio materials and searching based on speaker identity is a growing interest. Works related with Speech Emotion Recognition [64] and Natural Language Processing (NLP) [65] are based in acoustic features from the speech.

The most common features [66] are Mel-Spectrograms, Wave frames, Spectrograms and MFCCs. All these features are based in the frequency domain analysis.

In this project we mainly rely on third-party APIs for speech recognition modules.

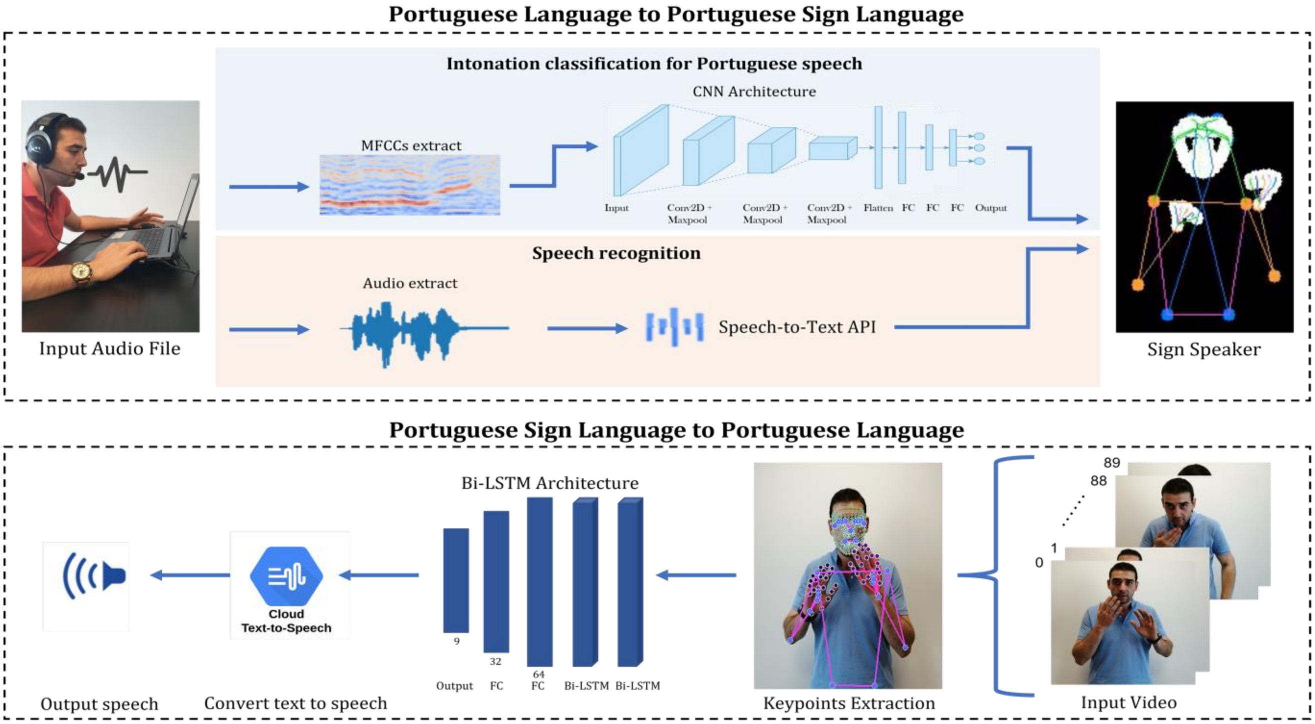

Figure 2: Architecture of WBG system. On top, first, we record a person talking to extract speech and after convert to text, at the same time we extract MFCCs to classify intonation to understand what is the punctuation of the sentence that has been said. At the bottom, we extract all body keypoints during 90 frames, then this sequence serves as input to our Bi-LSTM and finally the gestures are classified to give a sentence and then converted to speech.

3. Methodology

In this section, we present a description of our approach and the methods that were used to achieve our idea, as well as the creation of the two datasets used to train the proposed models. The block diagram of this system is illustrated in **Figure 2.**

For LGP to LP (bottom of the figure), first, we acquire a sequence of 90 frames, then we use Mediapipe Holistic [56] to extract all body keypoints, in each frame. Next, the resulting keypoints serve as input for the proposed network which is Bi-LSTM, to give as an output one of nine possible sentences. After the proposed network gives the result, we convert the written sentence into a spoken sentence.

For LP to LGP (top of the figure), first, we record an audio file that contains the sentence that was said by the emergency line assistant, then we classify the sentence with type, such as, Interrogative, Declarative, or Exclamative. To perform this classification we extract MFCCs features [67], that serve as input for our 2D-CNN model. At the same time, we convert speech to text using the Speech-to-Text API. Finally, we search in our database for the correspondent video of the sentence, and then a skeleton will reproduce Portuguese Sign Language.

3.1. Portuguese Sign Language to Portuguese Language

3.1.1. Realtime 3D all body Estimation

Joints detection is an essential step in the proposed system to ensure efficient processing before of SLR step. To implement this step a 3D all body keypoint [56] released in 2020 was used. This detector already has support in devices with less capability, like smartphones, which is a major advantage to the future implementation of the proposed system in these devices. This detector is a junction of previous works, such as [68]–[70].

The resultant features of this 3D keypoint detector are: 468*3 facial landmarks, 21*3 (per hand) hand landmarks, and 33*4 pose landmarks. The x and y coordinates of all landmarks are normalized between 0 and 1 by the image width and height, respectively. The z coordinate represents the depth with the depth at the midpoint of hips being the origin, so, the smaller value closer the person is to the camera. For pose landmarks this model still has a fourth coordinate, which is visibility to indicate if the joint is present or occluded in the image.

3.1.2. Bidirectional Long Short-Term Memory

Since videos are sequences of images, to encode temporal changes, a recurrent neural network (RNN) is necessary. Owing to the recurrent connections of each unit, RNNs yield good performances in modeling hidden sequential patterns of data. However, the internal memory property and the vanishing gradient problem of the RNN structure make it difficult to update the network parameters during backpropagation. Fortunately, the LSTM model conquers this fundamental weakness through its unique cell structure, which contains forget, input and output gates controlled by sigmoid units to decide what information must be updated and stored in memory cells. The linear connections across these units help to transmit the previous information to the present time step [71]. To generate a higher global context of sequential data from videos, we decided to use a Bi-LSTM model. The main difference between Bi-LSTM (3) and LSTM (4) reveals itself in the output layer because in (4) there only exists a single standard layer in time t. Therefore in (3) the combined output layers are decided by the previous and upcoming vectors. In other words, by employing the Bi-LSTM network in our model, the long-term bidirectional global temporal relationships were abstracted by going forward and backward several times through the vector sequences encoded from all the video segments.

The proposed model is based on [72], which contains three types of layers. The first type is the input layer that receives a sequence of 90 frames containing all keypoints from the skeleton in the original video. The second type is a Bi-LSTM where each frame enters in a forward and backward layer with a hidden state of 256 nodes (first layer), and 128 nodes (second layer). The third type is a series of fully connected layers of size 64, 32, 26, and finally a binary output layer with a SoftMax activation function, to predict the sentence that was gestured. Each one of the fully connected layers uses a RELU activation. The inputs of this model have a specific format which is (samples, frames number, number of coordinates).

3.1.3. Text to Speech

Since the classification of our Bi-LSTM model is in text, and the idea of this system it's to be based only on audio communication, there is the need to convert text to speech. For this implementation we use [73], which is an online API, to write spoken audio data to a file, and then using [74] to reproduce this file with a female voice.

The offline performance is very important to avoid mobile signal failure, enabling these emergency lines to be accessible anywhere. Instead of proposed systems by [5], [6] that perform online, our system can perform offline if we replace [73] with [75], which is an offline library to the same effect.

3.2. Portuguese Language to Portuguese Sign Language

3.2.1. Intonation Classification

As mentioned above, there are different gestures to represent the same word or sentence, in this way and as we only work with full sentences, there was a need to classify the intonation of the speech. To classify intonation, we used the same strategy that is used in Speech Emotion Recognition, which is features extraction from voice.

For this task, using only one sound channel, first, we remove noise by applying a band-pass filter between 20 Hz and 3.6 KHz. These frequencies allow keeping the speech 100% clean and understandable [76]. Inspired on [77] we take in filtered audios and extract 128 Mel Frequency Cepstral Coefficients (MFCCs) features to distinguish the sentence type, such as Declarative, Interrogative and Exclamative. This distinction is necessary because the same sentence with different punctuation has distinct gestures. The resultant features from MFCCs have the shape of (samples, MFCCs features, timesteps).

Since the proposed model is a 2D-CNN, input data must be reshaped to (samples, MFCCs features, timesteps,1) to be sent to this model. The overall architecture of the 2D-CNN model is depicted on top of the **Figure 2,** which contains three convolution layers, three polled layers and four fully connected layers. We input the MFCCs features into the input layer. This layer uses 32 convolutional kernels with the size of 3*3 to extract features after we connect a maximum pooling layer with the pool size of 2*2 to reduce the number of parameters (the pool size is the same for all layers). The other convolutional layers use 64 convolutional kernels with the size of 3*3. The end of the network is the fully connected layer with three neurons, SoftMax function is used to complete intonation classification.

3.2.2. Speech to text

At the same time that the proposed 2D-CNN model predicts the sentence type, we use [78] which is an online API to convert speech to text. Similarly to the text to speech conversion, also this conversion can be performed offline if we replace [78] with [79].

After the model predicts the result and the speech is converted, we search our video-sentence database [80]. This video-sentence database is detailed in section 4.3. We search this database in two ways. First, we search for the sentence type and then we search for the sentence that has been said. From the video, the skeleton is extracted from every frame and its movement is reproduced.

4. Experiments

This section describes experiments regarding the LGP and the LP recognition. To evaluate the effectiveness of the proposed system, two new datasets were collected for audio and signs. All experiences were implemented with TensorFlow 2.6, PyTorch 1.9, Cuda 11.2 and Cudnn 8.2.2.

4.1. Portuguese Sign Language Dataset

Since it was not possible to obtain the dataset created by [21] because General Data Protection Regulation (GDPR) issues. Furthermore, it doesn't exist anymore containing videos, so a new dataset was created for this project. This new dataset contains 9 sentences, in which sentence has 1000 video examples are distributed by 11 non-LGP speakers. To record these examples, we used a laptop webcam at the resolution of 640*480 at 30 fps. The videos each have 90 frames (3 seconds).

4.2. Intonation Dataset

Normally, intonation classification is used to classify emotions [65]. Because of this there was a need to create a new dataset with specific labels. This dataset contains 3 classes, each class has 500 examples.

To create this dataset, 2 people (one male and one female), were recorded during four seconds at the sample rate of 44100 Hz saying the sentences corresponding to gestures in the video-sentence database. To record these people, a microphone laptop was used in stereo mode.

4.3. Video-sentence Database

To collect the existent sentences in this database, first, we used Web Scrapping techniques to obtain the existent sentences in [80], then we use the work from [81] to extract the skeleton of the person present in video. This database contains the original videos and the videos the skeleton extracted.

4.4. Experimental Settings

In this paper two different types of neural networks were implemented. One type is a recurrent neural network, and the other type is a convolutional neural network. To produce a model capable to recognize Portuguese Sign Language, a Bi-LSTM architecture was used. The 2D-CNN architecture was used to identify intonation. Both models were implemented using Tensorflow/Keras library.

The Bi-LSTM model receives as input 1662 coordinates (per frame) extracted from 90 frames in each video. These coordinates form a sequence of a one-dimensional array. The model uses categorical_cossentropy as the loss function and Adam as the optimizer with an initial learning rate of 0.00001. The learning rate is reduced by half after 25 epochs if no improvement in categorical_accuracy is achieved. The model was trained with 300 examples per class and validated with 100 examples per class during training. The model stops training if no improvement in categorical_accuracy is achieved after 40 epochs. We have also experimented with different types of RNNs for comparison.

From each audio file 128 MFCCs features are extracted, and then normalized to serve as input to the 2D-CNN model. To normalize data, we use mean and standard deviation. The model uses categorical_cossentropy as the loss function and Adam as the optimizer. The data was initially split into 80% for training and validation and 20% for test. The 80% data for training and validation was further split into 90% for training and 10% for validation. The model stops training if no improvement in val_categorical_accuracy is achieved after 20 epochs. We have also experimented different combinations of input features, such as, Tonnetz [82] and MFCCs. Both features were extracted using librosa library.

Experiment results are demonstrated in section 5.

5. Results & Discussion

In this section, we report the accuracy of the proposed models in different experiences with our datasets. At the end of this section, we discuss our main problems and ways to solve them.

5.1. Results

Inspired in [83], we decided to experiment four variants of RNNs. These types are SimpleRNN, GRU, LSTM, Bi-LSTM. All models were trained with the same data and the same nodes per layer. As can be seen in **Table I**, all variants obtain the same results in the test set. The main difference that we note, was verified in the number of epochs of training however, these results may change depending on the characteristics of the computer used.

Table I: Classification Results of different RNNs

Networks	Accuracy (%)	Precision (%)	Recall (%)	Epochs
Bi-LSTM	99	99	99	47
RNN	99	99	99	241
LSTM	99	99	99	110
GRU	99	99	99	188

Regarding the model used to classify intonation, we decided to experiment different features as input in the same model. The different features are 128 MFCCs, 128 MFCCs + Tonnetz, and Tonnetz. Although the vocal frequencies are different between the male and female existent in the Intonation Dataset, when they say sentences of different types, the harmonic changes are similar. For that reason, we decided to experiment Tonnetz as one of the used features. The results obtained in the test set can be seen in **Table II**.

Table II: Classification Results of the 2D-CNN with different features

Features	Accuracy (%)	Precision (%)	Recall (%)
MFCCs	88	88	88
Tonnetz	65	65	65
MFCCs + Tonnetz	87	87	87

When we look at this table is possible to verify that Tonnetz doesn't bring any advantage when added to the MFCCs features.

5.2. Discussion

During the development and evaluation of the proposed system, several limitations emerged. The main limitation was found in the number of existing data with the necessary features for this work. For that reason, two new datasets were created for this project.

To create LGP dataset, first, we tried to use videos with examples available online, however under the advice of a sign language teacher working with individual words isn't a correct approach due to the context in which they are inserted, or because joining two words results in only one gesture. In this way, 9 different sentences were collected, contextualized with the emergency lines. The biggest disadvantage of the collected dataset is that it doesn't contain any LGP speakers.

Regarding the Intonation dataset, the main disadvantage is on the number of people present. Unfortunately, to create this dataset there weren't enough volunteers to read the sentences, so it not possible to have robustness in this dataset.

The detector used also proved to be challenging, as it proved to be sensitive to the textures of the clothes used by people during the recording of the videos, that is, it confuses these textures with possible keypoints.

6. Conclusion

In this paper, we bridge the gap between the deaf community and emergency lines assistants, applying deep learning techniques. The proposed system is capable of translating Portuguese Sign Language into spoken Portuguese Language, and vice-versa. Due to the problem faced by the avatars in reproducing eyes and facial movements, in this project we prefer not to introduce one in the system.

In the future work, we want to create a new detector to only extract arms, hands, and face keypoints, as applying artificial intelligence into an avatar to learn how to speak Sign Language correctly. We also intend to improve and expand the intonation dataset, video-sentence database, and recreate the Portuguese Sign Language dataset recording with LGP speakers.

References

[1] World Health Organization., "World report on Hearing," *World Rep. Hear.*, pp. 1–272, 2021.

[2] "Infográfico INSEF – Dificuldades auditivas - INSA." http://www.insa.min-saude.pt/infografico-insef-dificuldades-auditivas/ (accessed Aug. 27, 2021).

[3]. :: "GAF ::. Sítio Oficial do Gabinete de Atendimento à Família." https://www.gaf.pt/pt/noticias/sabia-que-a-lingua-gestual-portuguesa-lgp-e-uma-das-3-linguas-oficiais-em-portugal (accessed Aug. 27, 2021).

[4] "Há dois anos que surdos esperam por 'app 'que permite contactar 112 | Língua Gestual Portuguesa | PÚBLICO." https://www.publico.pt/2019/01/19/sociedade/noticia/governo-comprometeuse-ha-dois-anos-surdos-continuam-nao-conseguir-contactar-112-1858435 (accessed Aug. 27, 2021).

[5] "Secretaria Geral do MAI." https://www.sg.mai.gov.pt/tecnologias/112pt/appmobilepcidadaossurdos/Paginas/default.aspx (accessed Aug. 27, 2021).

[6] "Contacto acessível para cidadão surdo | SNS24." https://www.sns24.gov.pt/contacto-acessivel-cidadao-surdo/ (accessed Aug. 27, 2021).

[7] "Dicionário de língua gestual | SpreadTheSign." https://www.spreadthesign.com/pt.pt/search/ (accessed Aug. 28, 2021).

[8] "Serviin." http://www.portaldocidadaosurdo.pt/Serviin (accessed Aug. 28, 2021).

[9] Marcelo Norberto *et al.*, "Virtual Sign—Using a Bidirectional Translator in Serious Games," *China-USA Bus. Rev.*, vol. 14, no. 5, May 2015, doi: 10.17265/1537-1514/2015.05.004.

[10] "Início | CM Matosinhos." https://www.cm-matosinhos.pt/ (accessed Aug. 28, 2021).

[11] "CM Guimarães." https://www.cm-guimaraes.pt/ (accessed Aug. 28, 2021).

[12] "Avatar ajuda surdos a viajar na rede do Metro do Porto - Portal de notícias do Porto. Ponto." https://www.porto.pt/pt/noticia/avatar-ajuda-surdos-a-viajar-na-rede-do-metro-do-porto (accessed Aug. 28, 2021).

[13] "palexandrefernandes/P4-PT2LGP: Natural Portuguese Language to Portuguese Sign Language." https://github.com/palexandrefernandes/P4-PT2LGP (accessed Aug. 28, 2021).

[14] "ligar | Definição ou significado de ligar no Dicionário Infopédia de Língua Gestual Portuguesa." https://www.infopedia.pt/dicionarios/lingua-gestual/ligar (accessed Aug. 28, 2021).

[15] O. Roberto Santiago Amarante Oliveira, "TRADUTOR DA LINGUA GESTUAL PORTUGUESA MODELO DE TRADUÇÃO BIDIRECCIONAL."

[16] J. L. B. Lopes, "Tradutor bidirecional de língua gestual portuguesa," 2016, Accessed: Aug. 28, 2021. [Online]. Available: https://recipp.ipp.pt/handle/10400.22/11007.

[17] A. de O. M. C. Costa, "Reconhecimento de língua gestual," 2014, Accessed: Aug. 28, 2021. [Online]. Available: https://repositorio.ipl.pt/handle/10400.21/4329.

[18] I. G. S. Neiva, "Desenvolvimento de um tradutor de Língua Gestual Portuguesa," 2015, Accessed: Aug. 28, 2021. [Online]. Available: https://run.unl.pt/handle/10362/14753.

[19] S. C. M. Ferreira, "Sistema de tradução da língua gestual

Portuguesa em tempo real," Dec. 2019, Accessed: Aug. 27, 2021. [Online]. Available: https://repositorio.iscte-iul.pt/handle/10071/22215.

[20] P. M. M. Ferreira, "Sign Language Recognition: Integrating Prior Domain Knowledge into Deep Neural Networks," Mar. 2020, Accessed: Aug. 27, 2021. [Online]. Available: https://repositorio-aberto.up.pt/handle/10216/126550.

[21] and R. Ferreira, P. M., Rodrigues, I. V., Rio, A., Sousa, R., Pereira, E. M., "Corsil: A novel dataset for portuguese sign language and expressiveness recognition," 2014, pp. 1–2.

[22] P. M. Ferreira, D. Pernes, A. Rebelo, and J. S. Cardoso, "DeSIRe: Deep Signer-Invariant Representations for Sign Language Recognition," *IEEE Trans. Syst. Man, Cybern. Syst.*, vol. 51, no. 9, pp. 5830–5845, Dec. 2019, doi: 10.1109/TSMC.2019.2957347.

[23] R. Rastgoo, K. Kiani, and S. Escalera, "Sign Language Recognition: A Deep Survey," *Expert Syst. Appl.*, vol. 164, p. 113794, Feb. 2021, doi: 10.1016/J.ESWA.2020.113794.

[24] K. Grobel and M. Assan, "Isolated sign language recognition using Hidden Markov Models," *Proc. IEEE Int. Conf. Syst. Man Cybern.*, vol. 1, pp. 162–167, 1997, doi: 10.1109/ICSMC.1997.625742.

[25] V. Kulkarni, "Appearance Based Recognition of American Sign Language Using Gesture Segmentation," *undefined*, 2010.

[26] Z. Zafrulla, H. Brashear, T. Starner, H. Hamilton, and P. Presti, "American sign language recognition with the kinect," *ICMI'11 - Proc. 2011 ACM Int. Conf. Multimodal Interact.*, pp. 279–286, 2011, doi: 10.1145/2070481.2070532.

[27] J. Huang, W. Zhou, H. Li, and W. Li, "Sign Language Recognition using 3D convolutional neural networks," *Proc. - IEEE Int. Conf. Multimed. Expo*, vol. 2015-August, Aug. 2015, doi: 10.1109/ICME.2015.7177428.

[28] L. Pigou, M. Van Herreweghe, and J. Dambre, "Gesture and Sign Language Recognition with Temporal Residual Networks," *Proc. - 2017 IEEE Int. Conf. Comput. Vis. Work. ICCVW 2017*, vol. 2018-January, pp. 3086–3093, Jul. 2017, doi: 10.1109/ICCVW.2017.365.

[29] D. Metaxas, M. Dilsizian, and C. Neidle, "Scalable ASL sign recognition using model-based machine learning and linguistically annotated corpora," Feb. 2018, Accessed: Aug. 29, 2021. [Online]. Available: https://open.bu.edu/handle/2144/30049.

[30] Y. Quan, "Chinese sign language recognition based on video sequence appearance modeling," *Proc. 2010 5th IEEE Conf. Ind. Electron. Appl. ICIEA 2010*, pp. 1537–1542, 2010, doi: 10.1109/ICIEA.2010.5514688.

[31] F. Yasir, P. W. C. Prasad, A. Alsadoon, and A. Elchouemi, "SIFT based approach on Bangla sign language recognition," *2015 IEEE 8th Int. Work. Comput. Intell. Appl. IWCIA 2015 - Proc.*, pp. 35–39, Apr. 2016, doi: 10.1109/IWCIA.2015.7449458.

[32] A. Tharwat, T. Gaber, A. E. Hassanien, M. K. Shahin, and B. Refaat, "SIFT-Based Arabic Sign Language Recognition System," *Adv. Intell. Syst. Comput.*, vol. 334, pp. 359–370, 2015, doi: 10.1007/978-3-319-13572-4_30.

[33] H. Cooper, E.-J. Ong, N. Pugeault, and R. Bowden, "Sign Language Recognition Using Sub-units," *J. Mach. Learn. Res.*, vol. 13, pp. 89–118, Jul. 2017, doi: 10.1007/978-3-319-57021-1_3.

[34] P. Buehler, A. Zisserman, and M. Everingham, "Learning sign language by watching TV (using weakly aligned subtitles)," pp. 2961–2968, Mar. 2010, doi: 10.1109/CVPR.2009.5206523.

[35] S. Liwicki and M. Everingham, "Automatic recognition of fingerspelled words in British Sign Language," pp. 50–57, Mar. 2010, doi: 10.1109/CVPRW.2009.5204291.

[36] P. C. Badhe and V. Kulkarni, "Indian sign language translator using gesture recognition algorithm," *2015 IEEE Int. Conf. Comput. Graph. Vis. Inf. Secur. CGVIS 2015*, pp. 195–200, Apr. 2016, doi: 10.1109/CGVIS.2015.7449921.

[37] M. AL-Rousan, K. Assaleh, and A. Tala'a, "Video-based signer-independent Arabic sign language recognition using hidden Markov models," *Appl. Soft Comput.*, vol. 9, no. 3, pp. 990–999, Jun. 2009, doi: 10.1016/J.ASOC.2009.01.002.

[38] J. F. Lichtenauer, E. A. Hendriks, and M. J. T. Reinders, "Sign language recognition by combining statistical DTW and independent classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 11, pp. 2040–2046, 2008, doi: 10.1109/TPAMI.2008.123.

[39] S. Nagarajan and T. S. Subashini, "Static Hand Gesture Recognition for Sign Language Alphabets using Edge Oriented Histogram and Multi Class SVM," *Int. J. Comput. Appl.*, vol. 82, no. 4, pp. 28–35, Nov. 2013, doi: 10.5120/14106-2145.

[40] M. Borg and K. P. Camilleri, "Phonologically-Meaningful Subunits for Deep Learning-Based Sign Language Recognition," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12536 LNCS, pp. 199–217, 2020, doi: 10.1007/978-3-030-66096-3_15.

[41] A. Kratimenos, G. Pavlakos, and P. Maragos, "Independent Sign Language Recognition with 3D Body, Hands, and Face Reconstruction," Nov. 2020, Accessed: Aug. 29, 2021. [Online]. Available: http://arxiv.org/abs/2012.05698.

[42] X. Liang, A. Angelopoulou, E. Kapetanios, B. Woll, R. Al-batat, and T. Woolfe, "A Multi-modal Machine Learning Approach and Toolkit to Automate Recognition of Early Stages of Dementia among British Sign Language Users," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12536 LNCS, pp. 278–293, Oct. 2020, doi: 10.1007/978-3-030-66096-3_20.

[43] A. Moryossef, I. Tsochantaridis, R. Aharoni, S. Ebling, and S. Narayanan, "Real-Time Sign Language Detection using Human Pose Estimation," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12536 LNCS, pp. 237–248, Aug. 2020, Accessed: Aug. 29, 2021. [Online]. Available: https://arxiv.org/abs/2008.04637v2.

[44] R. Rastgoo, K. Kiani, and S. Escalera, "Real-time isolated hand sign language recognition using deep networks and SVD," *J. Ambient Intell. Humaniz.*

Comput., Feb. 2021, doi: 10.1007/S12652-021-02920-8.

[45] R. Rastgoo, K. Kiani, and S. Escalera, "Hand pose aware multimodal isolated sign language recognition," *Multimed. Tools Appl.*, vol. 80, no. 1, pp. 127–163, Jan. 2021, doi: 10.1007/S11042-020-09700-0.

[46] R. Rastgoo, K. Kiani, and S. Escalera, "Video-based isolated hand sign language recognition using a deep cascaded model," *Multimed. Tools Appl.*, vol. 79, no. 31–32, pp. 22965–22987, Aug. 2020, doi: 10.1007/S11042-020-09048-5.

[47] R. Rastgoo, K. Kiani, and S. Escalera, "Hand sign language recognition using multi-view hand skeleton," *Expert Syst. Appl.*, vol. 150, Jul. 2020, doi: 10.1016/J.ESWA.2020.113336.

[48] M. Parelli, K. Papadimitriou, G. Potamianos, G. Pavlakos, and P. Maragos, "Exploiting 3D Hand Pose Estimation in Deep Learning-Based Sign Language Recognition from RGB Videos," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12536 LNCS, pp. 249–263, Aug. 2020, doi: 10.1007/978-3-030-66096-3_18.

[49] R. Rastgoo, K. Kiani, and S. Escalera, "Multi-Modal Deep Hand Sign Language Recognition in Still Images Using Restricted Boltzmann Machine," *Entropy 2018, Vol. 20, Page 809*, vol. 20, no. 11, p. 809, Oct. 2018, doi: 10.3390/E20110809.

[50] S. N. Boualia and N. E. Ben Amara, "Pose-based human activity recognition: A review," *2019 15th Int. Wirel. Commun. Mob. Comput. Conf. IWCMC 2019*, pp. 1468–1475, Jun. 2019, doi: 10.1109/IWCMC.2019.8766694.

[51] D. C. Luvizon, D. Picard, and H. Tabia, "2D/3D Pose Estimation and Action Recognition using Multitask Deep Learning."

[52] A. Gupta, K. Gupta, K. Gupta, and K. Gupta, "Human Activity Recognition Using Pose Estimation and Machine Learning Algorithm."

[53] L. Song, G. Yu, J. Yuan, and Z. Liu, "Human pose estimation and its application to action recognition: A survey," *J. Vis. Commun. Image Represent.*, vol. 76, p. 103055, Apr. 2021, doi: 10.1016/J.JVCIR.2021.103055.

[54] Z. Cao, G. Hidalgo, T. Simon, S. E. Wei, and Y. Sheikh, "OpenPose: Realtime Multi-Person 2D Pose Estimation Using Part Affinity Fields," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 1, pp. 172–186, Jan. 2021, doi: 10.1109/TPAMI.2019.2929257.

[55] E. D'Antonio, J. Taborri, E. Palermo, S. Rossi, and F. Patane, "A markerless system for gait analysis based on OpenPose library," *I2MTC 2020 - Int. Instrum. Meas. Technol. Conf. Proc.*, May 2020, doi: 10.1109/I2MTC43012.2020.9128918.

[56] "Google AI Blog: MediaPipe Holistic — Simultaneous Face, Hand and Pose Prediction, on Device." https://ai.googleblog.com/2020/12/mediapipe-holistic-simultaneous-face.html (accessed Aug. 30, 2021).

[57] D. Osokin, "Real-time 2D Multi-Person Pose Estimation on CPU: Lightweight OpenPose," *ICPRAM 2019 - Proc. 8th Int. Conf. Pattern Recognit. Appl. Methods*, pp. 744–748, Nov. 2018, Accessed: Aug. 31, 2021. [Online]. Available: https://arxiv.org/abs/1811.12004v1.

[58] "Next-Generation Pose Detection with MoveNet and TensorFlow.js — The TensorFlow Blog." https://blog.tensorflow.org/2021/05/next-generation-pose-detection-with-movenet-and-tensorflowjs.html (accessed Aug. 31, 2021).

[59] P. Weinzaepfel, R. Brégier, H. Combaluzier, V. Leroy, and G. Rogez, "DOPE: Distillation of Part Experts for Whole-Body 3D Pose Estimation in the Wild," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 12371 LNCS, pp. 380–397, 2020, doi: 10.1007/978-3-030-58574-7_23.

[60] S. Stoll, N. C. Camgoz, S. Hadfield, and R. Bowden, "Text2Sign: Towards Sign Language Production Using Neural Machine Translation and Generative Adversarial Networks," *Int. J. Comput. Vis. 2019 1284*, vol. 128, no. 4, pp. 891–908, Jan. 2020, doi: 10.1007/S11263-019-01281-2.

[61] M. Kipp, A. Heloir, and Q. Nguyen, "Sign Language Avatars: Animation and Comprehensibility," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 6895 LNAI, pp. 113–126, 2011, doi: 10.1007/978-3-642-23974-8_13.

[62] S. Ebling and M. Huenerfauth, "Bridging the gap between sign language machine translation and sign language animation using sequence classification," pp. 2–9, Dec. 2015, doi: 10.18653/V1/W15-5102.

[63] S. Ebling and J. Glauert, "Exploiting the full potential of JASigning to build an avatar signing train announcements," *undefined*, 2013.

[64] Z. Peng, Y. Lu, S. Pan, and Y. Liu, "Efficient Speech Emotion Recognition Using Multi-Scale CNN and Attention," pp. 3020–3024, Jun. 2021, doi: 10.1109/icassp39728.2021.9414286.

[65] R. Ardila *et al.*, "Common voice: A massively-multilingual speech corpus," *Lr. 2020 - 12th Int. Conf. Lang. Resour. Eval. Conf. Proc.*, pp. 4218–4222, 2020.

[66] R. A. Solovyev *et al.*, "Deep Learning Approaches for Understanding Simple Speech Commands," *2020 IEEE 40th Int. Conf. Electron. Nanotechnology, ELNANO 2020 - Proc.*, pp. 688–693, Apr. 2020, doi: 10.1109/ELNANO50318.2020.9088863.

[67] S. Furui, "Cepstral Analysis Technique for Automatic Speaker Verification," *IEEE Trans. Acoust.*, vol. 29, no. 2, pp. 254–272, 1981, doi: 10.1109/TASSP.1981.1163530.

[68] "Pose - mediapipe." https://google.github.io/mediapipe/solutions/pose.html (accessed Aug. 30, 2021).

[69] F. Zhang *et al.*, "MediaPipe Hands: On-device Real-time Hand Tracking," Jun. 2020, Accessed: Aug. 30, 2021. [Online]. Available: https://arxiv.org/abs/2006.10214v1.

[70] "Google Developers Blog: MediaPipe 3D Face Transform." https://developers.googleblog.com/2020/09/mediapipe-3d-face-transform.html (accessed Aug. 30, 2021).

[71] X. Wu and Q. Ji, "TBRNet: Two-Stream BiLSTM Residual Network for Video Action Recognition," *Algorithms 2020, Vol. 13, Page 169*, vol. 13, no. 7, p. 169, Jul. 2020, doi: 10.3390/A13070169.

[72] "TechicalPaper_Human Activity Recognition.pdf - Google Drive."

https://drive.google.com/file/d/16rAiqm2Eby1k92UdBs8iEL_n9HIKFV-p/view (accessed Aug. 30, 2021).

[73] "gTTS — gTTS documentation." https://gtts.readthedocs.io/en/latest/ (accessed Aug. 30, 2021).

[74] "playsound · PyPI." https://pypi.org/project/playsound/ (accessed Aug. 30, 2021).

[75] "pyttsx3 - Text-to-speech x-platform — pyttsx3 2.6 documentation." https://pyttsx3.readthedocs.io/en/latest/ (accessed Aug. 30, 2021).

[76] "Facts about speech intelligibility: human voice frequency range." https://www.dpamicrophones.com/mic-university/facts-about-speech-intelligibility (accessed Aug. 31, 2021).

[77] K. Venkataramanan and H. R. Rajamohan, "Emotion Recognition from Speech," *SpringerBriefs Speech Technol.*, pp. 31–32, Dec. 2019, Accessed: Aug. 30, 2021. [Online]. Available: https://arxiv.org/abs/1912.10458v1.

[78] "SpeechRecognition · PyPI." https://pypi.org/project/SpeechRecognition/ (accessed Aug. 31, 2021).

[79] "CMUSphinx Documentation – CMUSphinx Open Source Speech Recognition." https://cmusphinx.github.io/wiki/ (accessed Aug. 31, 2021).

[80] "Dicionário de língua gestual | SpreadTheSign." https://www.spreadthesign.com/pt.pt/search/ (accessed Sep. 03, 2021).

[81] "open-mmlab/mmpose: OpenMMLab Pose Estimation Toolbox and Benchmark." https://github.com/open-mmlab/mmpose (accessed Sep. 03, 2021).

[82] C. Harte, M. Sandler, and M. Gasser, "Detecting Harmonic Change In Musical Audio," *Proc. 1st ACM Work. Audio Music Comput. Multimed. - AMCMM '06*, 2006, doi: 10.1145/1178723.

[83] S. Yang, X. Yu, and Y. Zhou, "LSTM and GRU Neural Network Performance Comparison Study: Taking Yelp Review Dataset as an Example," *Proc. - 2020 Int. Work. Electron. Commun. Artif. Intell. IWECAI 2020*, pp. 98–101, Jun. 2020, doi: 10.1109/IWECAI50956.2020.00027.