# An Intelligent System to Automatically Generate Video-Summaries for Accessible Learning Objects for People with Hearing Loss

Paola Ingavélez-Guerra[1,2(✉)], Irma Cuzco-Calle[1], Daniel Calle-López[1], Christian Oyola-Flores[1], Inés Yambay-Aulla[1,2], Vladimir Robles-Bykbaev[1], and José Ramón Hilera[2]

[1] GI-IATa, Cátedra UNESCO Tecnologías de Apoyo a La Inclusión Educativa, Universidad Politécnica Salesiana, Cuenca, Ecuador
{pcingavelez, vrobles}@ups.edu.ec, {icuzco, dcallel, coyola, iyambay}@est.ups.edu.ec

[2] Universidad de Alcalá de Henares, Madrid, Spain
jose.hilera@uah.es

**Abstract.** The latest estimates of the World Health Organization point that approximately 5% of world's population presents disabling hearing loss (328 million adults and 32 million children). This situation becomes very complex in developing countries, where children with hearing loss and deafness rarely have access to schooling. On those grounds, in this paper we present an intelligent system to automatically generating video-summaries and captioning in sign language with the aim of creating accessible learning objects for children and youth with disabling hearing loss. Through our intelligent system, we propose a methodology that allows performing agile consumption of educational content (Learning Accessible Objects) that use videos. With the aim of determining the real feasibility of our proposal, we have performed a preliminary experiment with 7 educational videos of history, natural sciences and mathematics for deaf children and youth from 8 to 12 years. A team of 6 experts in sign language has evaluated several aspects our proposal (coherence of the summaries and captions, quality of the synchronization between the explicative elements in sign language and video execution, possible lost keywords/meanings, …).

**Keywords:** Hearing loss · Sign language · Deaf persons · Video summary · Children with disabilities · Learning objects

## 1 Introduction

The latest estimations of the World Health Organization point that nowadays more than 5% of world population present hearing loss. This percentage represents 360 million of persons, of whom 32 million are children [1]. Commonly, hearing loss is related with hereditary characteristics, perinatal trauma, infectious diseases, exposure to excessive noise, aging, the use of medication that produces side effects, among others causes.

In this line, the early detection of hearing problems as well as using cochlear implants and hearing aids allows incrementing opportunities for coexistence for persons


T. Andre (ed.), *Advances in Human Factors in Training, Education, and Learning Sciences*, Advances in Intelligent Systems and Computing 596,
DOI 10.1007/978-3-319-60018-5_12

with hearing loss. For a person with hearing loss it is fundamental to have the opportunity of using subtitles, sign language, lip-reading and other educational and social support strategies that constitute the mainstays of inclusion. Nowadays the Word Federation of the Deaf (WFD), one of leading worldwide non-governmental associations, works to promote the creation of opportunities (working, education, etc.) for people with different degrees of hearing loss. Similarly, the WFD promotes the correct use of sign language has a means of access to education, information and several aspects of common life. This organization actually represents to 70 million persons worldwide, grouping national organizations of 131 countries [2].

Another important aspect to consider is that interpreting sign language constitutes a diverse work of linguistic, social and cultural participation that involves special methods for receipting, processing and transmitting information. Nowadays, the most of countries must to face the challenge of providing a quality education for all persons, strengthening an approach of inclusion that gradually gains ground in educational and social areas to address the existing high levels of exclusion, discrimination and educational inequality. In order to guarantee educational equity, it is necessary to apply important changes in the educational systems, as well as the culture, policies and practices of governments, including for it evaluative processes that help validating the made efforts.

For these reasons, it must be taken into account that deaf persons have several problems to access to information contained in different media, especially in Internet. The continuous use of infrequent terms that appear in language as well as a complex/broad syntax makes very difficult for a deaf person understanding several concepts, ideas or messages. To make accessible multimedia materials, it is necessary applying adaptations based on subtitles, using images/diagrams and/or including sign language videos [3].

On those grounds, in this paper we present a complete ecosystem to automatic generating summaries of videos to create accessible learning objects for persons with different degrees of hearing loss. In the same way, we present the first experimentation stage as well as the results obtained in a center of special education in Cuenca, Ecuador.

The rest of the paper is organized as follows. In Sect. 2 we present some relevant contributions focused on providing support for educational contents and sign language. The system architecture as well as the study that we have carried out in Cuenca, Ecuador are depicted in Sect. 3. Section 4 describes the main findings and results of our research. Finally, Sect. 5 presents some ideas for discussion and future work.

## 2 The Sign Language and ICTs Tools: A Brief Overview

In the last years have been developed several contributions that are focused on developing intelligent systems able to recognize the sign language used by persons as well as creating virtual avatars that can maintain conversations with deaf people.

Dorfman et al. proposes a virtual agent able to interact with humans through natural language. This agent can simulate human dialogs with the objective of been a complement higher education processes. To this aim, the authors have used Natural

Language Processing (NLP) and conducted an experiment during two years in the IT Resource Management course at Buenos Aires University. The results of the 604 conversations collected during two years period show a great acceptance level by student's side [4]. In a similar research line, Uluer, Akalın, and Köse have presented a system assisted by a humanoid robot to teach sign language to children. This robot named Robovie R3 has five fingers and uses games based on imitation to teach children Turkish sign language. Robovie R3 is able to combine the gestures produced by its hands with facial expressions and body movements. Another feature of the robot is that it can recognize signs through its RGB-D video camera [5].

Gamage et al. have proposed using Gaussian process dynamical models for hand gesture interpretation in sign language as an alternative to Hidden Markov Models (HMM) or Artificial Neural Networks (ANN). This proposal has tested with a database of 66 hand gestures from the Malaysian Sign Language, obtaining a rate of 79% success. The authors conclude that their contribution is stronger in areas such as rigour in training (number of samples, parameters, and training time), over-learning tendencies and variance in classification [6]. Similarly, Huang et al. have developed a method for hand gestures interpretation through Real-Sense device, which includes a video camera able to detect and track the hands position. This device provides a set of 3D coordinate of fingers' articulations [7].

In the other hand, Centelles et al. have carried out an experiment in which the authors have assigned accessible metadata to educational videos at higher education context. This research uses Schema.org vocabulary in combination with three of the most used search engines (Google, Yahoo and Bing). In this first stage, the authors have established the pillars of a system that will be developed later and will provide rich snippets [8].

In the same way, several authors have developed systems and algorithms with the aim of performing video summarization [9–11]. However, this task is not trivial, given that the summaries are subjective: different persons can compose different summaries for the same video. As such, a video summary becomes very important for a deaf persons, because it can help them searching and finding the video content according to its semantics [9]. Creating video summaries for persons with hearing loss require an extra effort, given that is necessary carrying out an adequate "translation" from spoken language to sign language. This is a complex process in which it is necessary not only matching words with signs, but also interpret the message as it is done by a human expert.

## 3 General System Architecture

As can be seen in Fig. 1, our ecosystem is organized in 5 layers and several modules/components. This approach allows us performing changes on any module as well as testing new components without affecting the general functionality of the entire system. Below we describe the different functionalities and elements of the ecosystem:

- The system uses the dictionary provided by the **Ecuadorian National Council for Equality of Disabilities** (Consejo Nacional para la Igualdad de las Disapacidades,

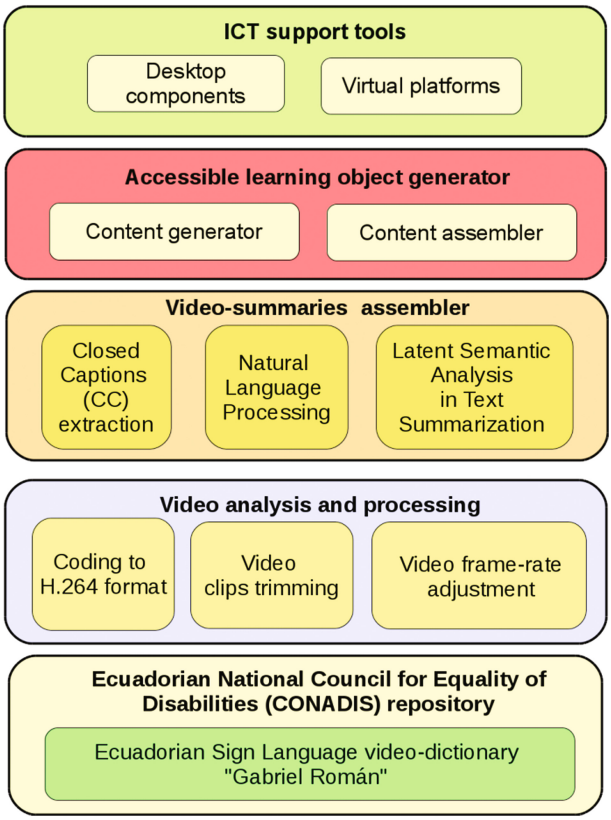


**Fig. 1.** The general system architecture and the 5 layers and components that constitute it.

CONADIS). The dictionary of Ecuadorian sign language "**Gabriel Román**" has approximately 5,000 words from the Official Dictionary of the Language of Ecuadorian Signs and includes graphics and explanatory videos. To carry out this task, the system performs a **video analysis and processing** stage that consists on the following activities:

(a) **Coding to H.264 format**: each video is analyzed and codified using the **H.264 format** (in order to use less space and maintain the quality),
(b) **Video clips trimming**: the "blank spaces" that do not contain information are removed (this allows reducing the video length from 5 to 2 s),
(c) **Video frame-rate adjustment**: the system reduces the frame-rate to 15 frames per second (this allows playing videos in mobile devices with reduced computational resources).

- The video-summaries assembler layer is in charge of generating the subtitles (captioning) and from this result assembles each summaries. To achieve this objective, this layer performs the following operations:

(a) **Closed Captions (CC) extraction**: The system uses the YouTube tools to access to video subtitles. From this information, the extractor creates several data structures that contain the texts and data-time (staring time and ending time) of each sentence.

(b) **Natural Language Processing (NLP)**: At this stage, the system performs a Part-Of-Speech tagging (POS tagging) with the aim of determining the most relevant words (verbs, pronouns, nouns, etc.). Each of these words will pair with a video of the Ecuadorian sign language dictionary. However, several texts contain words that are not registered in the dictionary, for these cases, the system searches an "equivalent" word through synonyms and Second-order co-occurrence pointwise mutual information method [12].

- **Latent Semantic Analysis (LSA) in Text Summarization**: At this stage, the system creates a new accessible LO that can be used by any person with hearing loss condition. The content generator uses the different videos created in the video analysis and processing layer to create new subtitles based not only on texts, but also on sequences of images with hand gestures representing signs (Fig. 2). These contents are used by content assembler to create the LO that will be charged to virtual learning environment (in this case MOODLE [13]).
- In order for providing users an interface to interact with the ecosystem, the ICT support tools layer provides two elements: a set of Desktop components and a set of virtual platforms (different kinds of applications to learn sign language through games). The desktop components are used as part of educational programs for children and youth whereas the virtual components are used as reinforcement activities for home.

In Fig. 2 is depicted the general process that we have followed to automatically generate the video captioning in Ecuadorian sign language. Below we describe the most relevant tasks (that were not described above):

- The first step consists on extracting the **Closed Captions (CC)** from the educational videos. This task is carried out with the support of **Google2SRT** tool. However, the sentences retrieved from Google2SRT are fragmented (commonly the timestamps appear at the middle of sentence). Therefore, to determine the beginning and ending of each sentence as well as joining the fragment, the system uses **regular expressions**.
- The system extracts from sentences that have time stamp the most relevant words (nouns, pronouns, verbs, etc.) through a lexical analysis (**PoS Tagging** and **Stop-words removal**). In some cases, it is necessary to search a synonym of some words, given that the Ecuadorian sign language only has 5,000 entries. For this task, the system uses **DISCO** and **Vector Space Model** (VSM) [12] to find the words semantically similar to the searched word.
- In order to generate the video summaries, the system applies **LSA** [14] to paragraphs assembled from the sentences in which were removed the time stamps. This summary presents the key ideas of the video in sign language, and can be used by user to search videos according to its content.

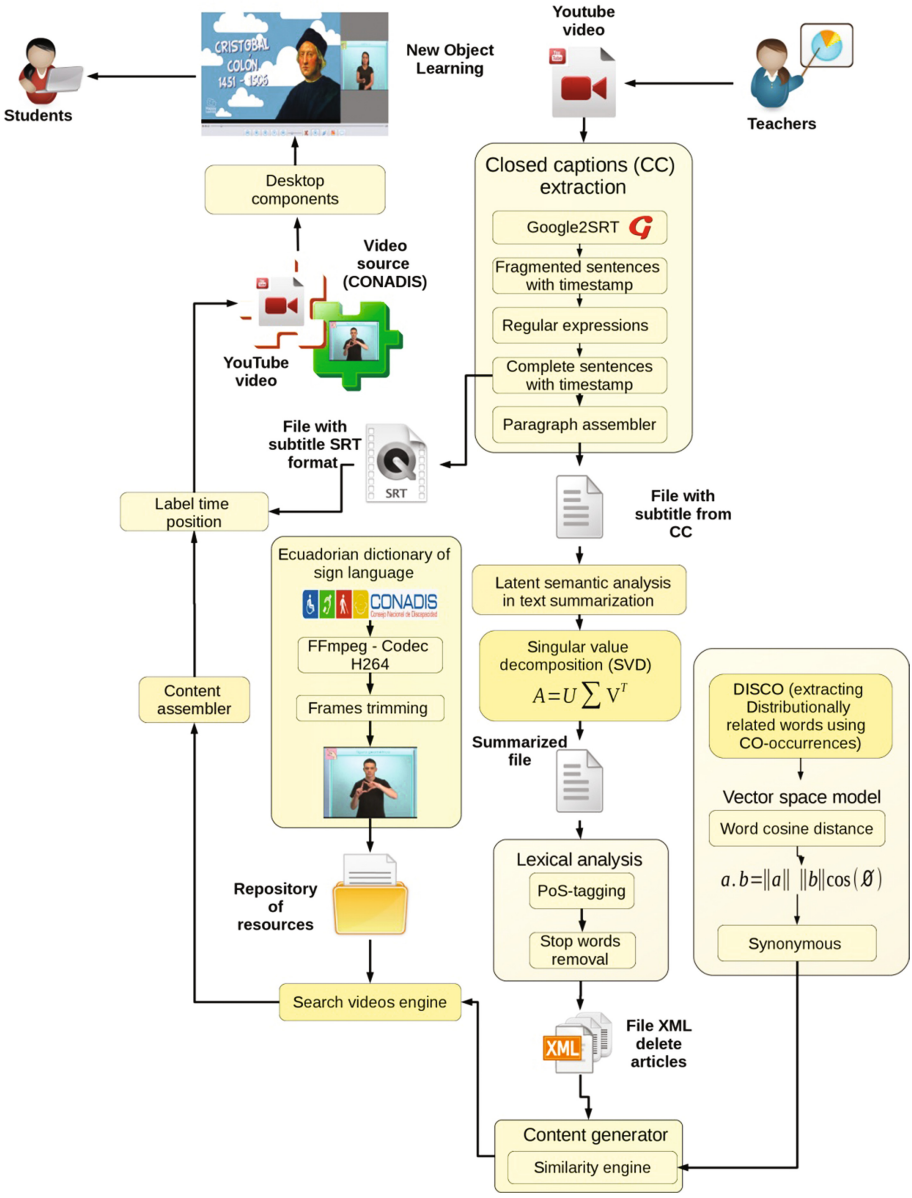


**Fig. 2.** Detailed diagram of the different tasks carried out by the system to generate both video captions and video summaries in Ecuadorian sign language.

In Fig. 3 it is possible to see a screen capture of main window that belongs to desktop application. On the left side we can see the standard educational video, whereas on the right side the system includes the sign language captions (animation) that are synchronized with the video-playing.

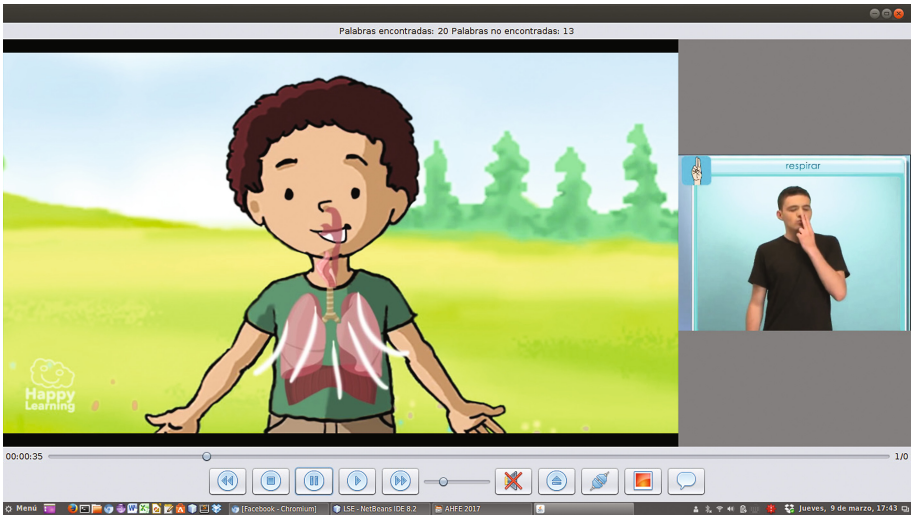


**Fig. 3.** A screen capture of the main window of user interface.

## 4 Pilot Experiment and Preliminary Results

With the aim of validating the first stage of our proposal, we have carried out a pilot experiment with a team of 6 experts in Ecuadorian sign language (interpreters). These experts collaborate with a public education institution where 11 children with hearing loss have been integrated.

The experiment has consisted in two parts, a first with the objective of determining key parameters to evaluate the developed tool. These parameters have the objective of measuring the relevance and real utility of the tool in educational area, and are the following:

- *Instruct*: specifies if the tool can be used in educational context to teach different courses aimed at children.
- *Inform*: considers whether the tool is appropriate to provide informative contents to users.
- *Motivate*: this parameter is used to measure the level of motivation with which users interact with the educational videos provided by the system.
- *Explore*: this parameter specifies whether the tool incorporates services to search information in the videos (according to certain criteria and topics).
- *Develop skills*: allows determining whether the tool can be used to develop some communication skills (based on sign language).
- *Entertain*: specifies whether is possible entertaining children and youth with the platform.
- *Experiment*: determines if the tool allows users testing functionalities related with labeling process as well as with summaries generation (in sign language).

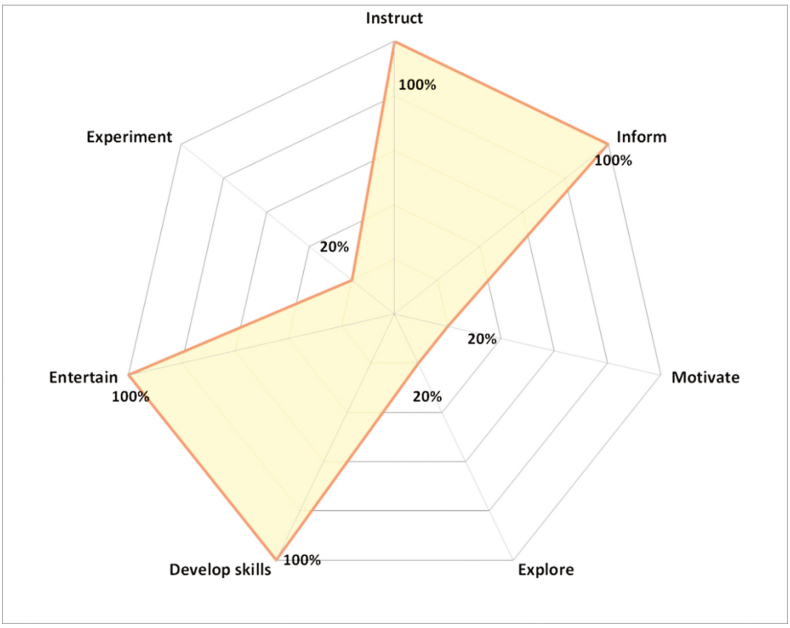


**Fig. 4.** Set of evaluation parameters defined by the expert team on Ecuadorian sign language.

As can be seen in Fig. 4, our proposal was evaluated considering the 7 parameters described above. To this aim, the experts have reproduced several educational videos and then they did respond a survey. The results show that we must improve 3 key areas: (a) the motivation of users to use the platform, (b) the alternatives to accomplish exploratory activities on the educational contents, and (c) the support to experiment with the contents.

As for the other parameters, it can be observed that although this tool constitutes a prototype, it is adequate to instruct, inform, support skills development, and entertain persons while they are learning.

**Table 1.** Evaluation results of the educational videos and their captions in Ecuadorian sign language.

| Video | Quality | Coherence | Synchronization |
|---|---|---|---|
| 1 | 5 | 3 | 4 |
| 2 | 4 | 3 | 4 |
| 3 | 4 | 3 | 3 |
| 4 | 3 | 2 | 3 |
| 5 | 5 | 4 | 4 |
| 6 | 4 | 3 | 4 |
| 7 | 5 | 4 | 4 |
| Average | 4.3 | 3.1 | 3.7 |

In the second part of the pilot experiment, the team of experts has evaluated the quality (the signs sequence is properly assembled), coherence (the message in sign language and spoken language are closely related) and synchronization (the captions in sign language are showing at correct times) of the captions in sign language. In Table 1 we can see the evaluation results for each of the three variables mentioned above. As it can be seen, it is necessary improving the coherence and synchronization of the captions in sign language.

## 5 Conclusions and Future Work

Nowadays, the demand for an inclusive education makes it necessary to have accessible and universal environments that allow access to all people, regardless of whether or not they have disabilities.

Therefore, it is important to mention that videos are very valuable learning objects, so giving options that allow synchronous or asynchronous communication with people with hearing impairment will strengthen the collective construction of knowledge.

It is necessary to propose a next phase focused on the semantic analysis of the deaf people's grammatical structure, since the interpretation goes beyond a translation, it requires considering 3 components associated with facial expression, body movement and the sign itself, for which we require analyzing patterns of behavior and generate an avatar that conveys the complete communication.

Through the use of interaction technologies such as Leap Motion and Kinect, we can properly control an avatar in order to generate expressions related to the Ecuadorian sign language. Likewise, this process will enrich the video captioning.

**Acknowledgements.** This work was funded by the Cátedra UNESCO Tecnologías de Apoyo para la Inclusión Educativa of the Universidad Politécnica Salesiana.

## References

1. World Health Organization (WHO): Deafness and hearing loss (2017)
2. World Federation of the Deaf (WFD). https://wfdeaf.org/
3. Debevc, M., Milošević, D., Kožuh, I.: A comparison of comprehension processes in sign language interpreter videos with or without captions. PloS One **10** (2015)
4. Dorfman, M., Grondona, A., Mazza, N., Mazza, P.: Asistentes Virtuales de Clase como complemento a la educación universitaria presencial. In: SADIO-40 JAIIO (2011)
5. Uluer, P., Akalın, N., Köse, H.: A new robotic platform for sign language tutoring. Int. J. Soc. Robot. **7**, 571–585 (2015)
6. Gamage, N., Kuang, Y.C., Akmeliawati, R., Demidenko, S.: Gaussian process dynamical models for hand gesture interpretation in sign language. Pattern Recognit. Lett. **32**, 2009–2014 (2011)
7. Huang, J., Zhou, W., Li, H., Li, W.: Sign language recognition using real-sense. In: 2015 IEEE China Summit and International Conference on Signal and Information Processing (ChinaSIP), pp. 166–170 (2015)

8. Centelles Velilla, M., Vázquez Guzmán, C., Ribera, M., Pérez Pineda, I.: Asignación de metadatos de accesibilidad a vídeos docentes (2016)
9. Demirtas, K., Cicekli, I., Cicekli, N. K: Summarization of documentaries. In: Computer and Information Sciences, pp. 105–108 (2011)
10. Potapov, D., Douze, M., Harchaoui, Z., Schmid, C.: Category-specific video summarization. In: European Conference on Computer Vision, pp. 540–555 (2014)
11. Xu, J., Mukherjee, L., Li, Y., Warner, J., Rehg, J. M., Singh, V.: Gaze-enabled egocentric video summarization via constrained submodular maximization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2235–2244 (2015)
12. Kolb, P.: Experiments on the difference between semantic similarity and relatedness. In: Proceedings of the Ordic Conference on Computational Linguistics (ODALIDA), pp. 81–88 (2009)
13. Büchner, A.: Moodle 3 Administration. Packt Publishing Ltd, Birmingham (2016)
14. Guerreiro, J., Gonçalves, D., De Matos, D.M.: Towards a fair comparison between name disambiguation approaches. In: Proceedings of the 10th Conference on Open Research Areas in Information Retrieval, pp. 17–20 (2013)