An on-line system adding subtitles and sign language to Spanish audio-visual content

Jordi Porta, Fernando López-Colino, Javier Tejedor, and José Colás

Human Computer Technology Laboratory,
Universidad Autónoma de Madrid, Spain,
jordi.porta@uam.es

Abstract. Deaf people cannot properly access the speech information stored in any kind of recording format (audio, video, etc). We present a system that provides with subtitling and Spanish Sign Language representation capabilities to allow Spanish Deaf population can access to such speech content. The system is composed by a speech recognition module, a machine translation module from Spanish to Spanish Sign Language and a Spanish Sign Language synthesis module. On the deaf person side, a user-friendly interface with subtitle and avatar components allows him/her to access the speech information.

1 Introduction

Spanish Sign Language (LSE) has been considered an official language in Spain since 2007. There are at about one million of Deaf people in Spain and at about 100 000 LSE users. In addition, the ratio between official LSE interpreters and LSE users is 1/221, which falls below the European average of other sign languages.

On the other hand, it has been widely accepted that LSE provides an efficient way for Spanish Deaf people to access to any kind of information, especially speech-based information, no matter the way it is stored (audio or video format) [4–6, 12]. Moreover, the increase in the amount of information stored in a speech-based format makes necessary the quick development of systems and services that allow Deaf people to access to it. To solve this issue, both subtitling and sign language-based systems can be effectively employed [4, 7, 8, 10].

We present an on-line system that produces both subtitles and Spanish Sign Language content aiming at translating the speech content of a video (e.g., TV news, meetings, weather forecast) so that a Spanish Deaf person can access to such content. In so doing, our demo will consist of the subtitled and signed of a Spanish speech content recorded previously in a video format.

Our system takes the speech of the video and sends it towards a speech recognition module that transcribes it into a sequence of words; next, a machine translation module translates this sequence of words into a sequence of glosses; and finally, a Sign Language synthesis module takes this sequence of glosses and generates an animation using an avatar. Both subtitle and avatar components have been included within a user-friendly interface that shows both the subtitles and the LSE content.

2 System description

The system architecture is depicted in Figure 1. The input towards the system is the speech corresponding to the selected video and the output consists of both the word transcription (i.e., subtitles) and the Spanish Sign Language representation of the speech. Our system is divided in the following modules: speech recognition, machine translation and sign language synthesis, which are explained next along with the user interface and middleware.

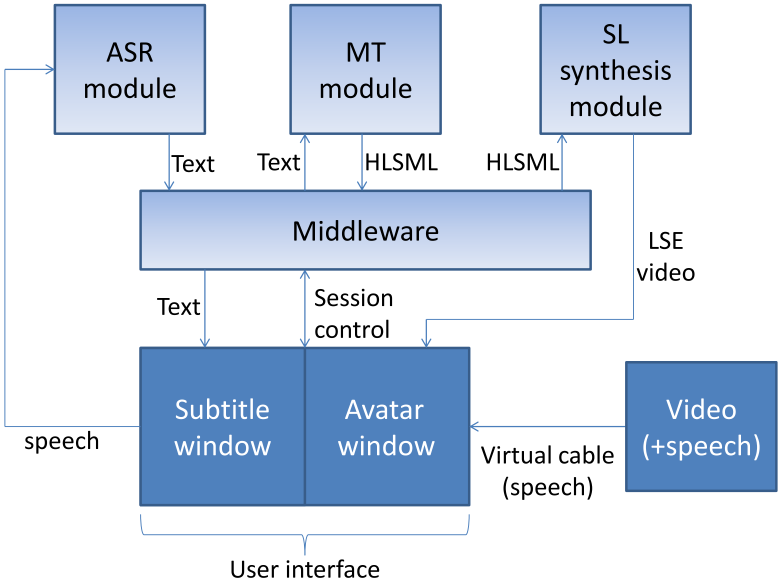

Fig. 1. *The system architecture. 'ASR' stands for Automatic Speech Recognition, 'MT' for Machine Translation, 'SL' for Sign Language and 'HLSML' for High Level Sign Mark-up Language.*

2.1 Speech recognition module

Speech is transcribed by means of the *ATK* software (an HTK-based API for on-line recognition) [14]. As acoustic modeling, and due to the limited training data available, standard 3-state context-independent left-to-right topology Hidden Markov Models were employed for speech decoding. Each state is modeled with 60 Gaussian Mixture Model components. These acoustic models were trained from the ALBAYZIN database [9] for a set of 47 phones [11] in Spanish plus beginning and end silences with 39-dimensional Mel Frequency Cepstrum Coefficient features. There is also a short pause model that contains a single emitting state and a skip transition. This module also employs a bigram language model during the speech decoding, which was trained from a corpus related to the disability domain, chosen for this demo, and a vocabulary that consists of 2000 words.

2.2 Machine translation module

The machine translation module performs a symbolic translation of a Spanish text into a sequence of glosses representing base signs with morphological annotations. The module has been designed following a transfer-based approach at the level of grammatical functions. The three main phases in which translation can be decomposed are: analysis, transfer and generation. In the analysis phase, the Spanish text is segmented into sentences, words, dates, proper nouns, quantities, etc., using Freeling [1]. Each sentence is parsed using a Spanish grammar with a relatively wide grammatical coverage, which copes with complementation, adjunction, pronominalization, extraction, etc. The type of grammar used is a constraint-based unification grammar with weights expressing preferences. The parser combines a weighted chart with an A* search strategy limited by a beam. In case of failure, the parser applies the shortest-path algorithm in order to recover good partial analyses from the weighted arcs of the chart. The parser produces a constituency tree from which a dependency tree is lift-up. During the transfer phase, the dependency tree is transferred to another LSE dependency tree using a transfer lexicon and some rules for inserting and removing nodes. For example, articles, prepositions and some pronouns are removed, and non-overt subjects are inserted. The transfer lexicon is based on the *Diccionario Normativo de la LSE* [2]. In order to extend the coverage of the lexicon, the morpho-lexical and lexico-semantic relations between words are exploited by using the Wordnet ontology [13]. Finger-spelling is also used in case a Spanish word does not have a sign representation. Generation phase can be subdivided into two subphases. In the first subphase, the dependency tree in LSE is traversed applying linear precedence rules to obtain the surface order of signs. Finally, signs with annotated morphological information are produced in a format according to the input notation (HLSML) of the Sign Language synthesis module.

2.3 Sign Language synthesis module

The Sign Language synthesis module presented in a previous work [3] has been used to synthesize the signs. It receives the message translated by the machine translation module in an HLSML-based format and parses it to produce the sequence of signs in the message. It employs a relational database that stores 500 different signs with their phonetic definition. This phonetic definition is then used to generate the corresponding animation sequence of the whole sentence by means of an avatar. This avatar animation is displayed using a real time rendering API to generate the corresponding video.

2.4 User interface and middleware

A friendly user interface is employed as input/output to/from the system. It sends the speech corresponding to the video to the speech recognition module and gets the video generated by the synthesis module by means of standard RTP connections. The speech is managed through a virtual cable that allows the video display software to send the speech stream to the output side of the cable so that the user interface can receive this stream from the virtual cable input side (as if it was a *real* sound card) and can send it towards the speech recognition module. The subtitle module of the system has been also included in this user interface and hence, this also receives the sequence of words output by the speech recognition module by means of a TCP/IP connection. Therefore, this interface, whose screenshot is depicted in Figure 2, consists of a subtitle window to show the word transcription output by the speech recognition module, and a window to show the signing avatar as main components.

The middleware manages the connection and the dataflow between the modules of the system from standard TCP/IP connections and between the speech recognition module and the subtitle module. Therefore, it sends the speech transcriptions from the speech recognition module to the machine translation module and the subtitle module, and the sequence of glosses from the machine translation module to the Sign Language synthesis module. Thus, the final user (i.e., the deaf person) only needs to install the user interface along with the virtual cable in his/her computer, apart from the appropriate video display software to play the corresponding video.

3 Conclusions

We present an on-line system with subtitling and Spanish Sign Language representation capabilities aiming at representing Spanish speech content. This will allow Spanish Deaf people to access speech information.

Future work will study new application domains such as TV news, weather forecast and so on. This requires extending the vocabulary coverage of the speech recognition and machine translation modules and adding the corresponding signs to the LSE relational database.

References

1. Atserias, J., Casas, B., Comelles, E., González, M., Padró, L., Padró, M.: FreeLing 1.3: Syntactic and semantic services in an open-source NLP library. In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006). pp. 48–55 (2006)
2. Fundación CNSE: Diccionario normativo de la lengua de signos española. Fundación CNSE (2008)
3. López-Colino, F., Colás, J.: Spanish sign language synthesis system. Journal of Visual Languages & Computing 23(3), 121–136 (2012)

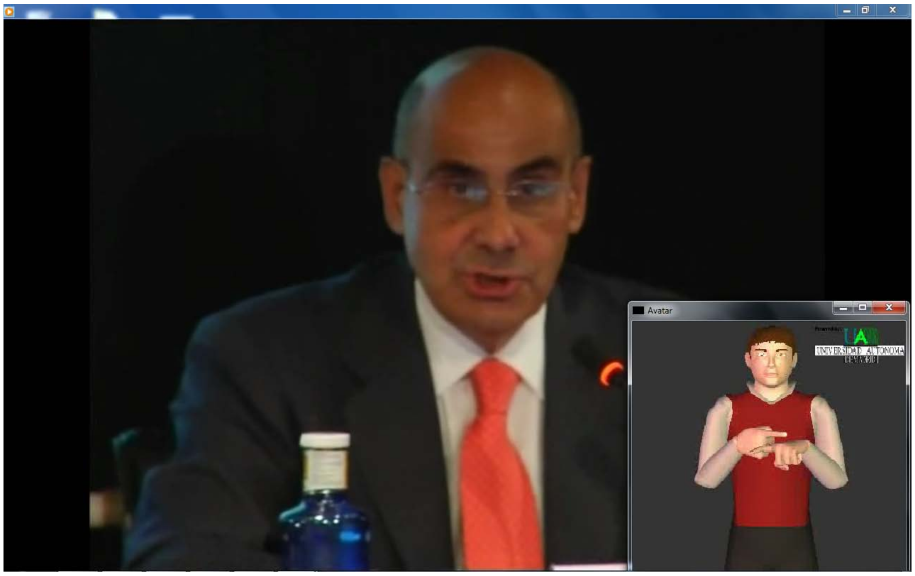

Fig. 2. *The user interface and the corresponding video. To avoid cluttering, the subtitle component is not shown in this screenshot.*

4. López-Colino, F., Tejedor, J., Porta, J., Colás, J.: Integration of a spanish-to-LSE machine translation system into an E-learning platform. In: Proceedings of the 6th international conference on Universal access in human-computer interaction: applications and services - Volume Part IV. pp. 567–576 (2011)
5. López-Ludeña, V., San-Segundo, R., de Córdoba, R., Ferreiros, J., Montero, J.M., Pardo, J.M.: Factored translation models for improving a speech into sign language translation system. In: Proceedings of Interspeech. pp. 1605–1608 (2011)
6. López-Ludeña, V., San-Segundo, R., Lutfi, S., Lucas-Cuesta, J., Echevarry, J., Martínez-González, B.: Source language categorization for improving a speech into sign language translation system. In: Proceedings of the Workshop on Speech and Language Processing for Assistive Technologies (SLPAT-2011). pp. 84–93 (2011)
7. López-Ludeña, V., San-Segundo, R., Martín, R., Sánchez, D., García, A.: Evaluating a speech communication system for deaf people. IEEE Latin America Transactions 9(4), 565–570 (2011)
8. Meinedo, H., Abad, A., Pellegrini, T., Neto, J., Trancoso, I.: The L2F broadcast news speech recognition system. In: Proceedings of FALA. pp. 93–96 (2010)
9. Moreno, A., Poch, D., Bonafonte, A., Lleida, E., Llisterri, J., Mariño, J., Nadeu, C.: Albayzin speech database: Design of the phonetic corpus. In: Proceedings of Eurospeech. pp. 653–656 (1993)
10. Ortega, A., García, J.E., Miguel, A., Lleida, E.: Real-time live broadcast news subtitling system for spanish. In: Proceedings of Interspeech. pp. 2095–2098 (2009)
11. Quilis, A.: El comentario fonológico y fonético de textos. ARCO/LIBROS, S.A. (1998)
12. San-Segundo, R., Barra, R., D'Haro, L.F., Montero, J.M., Córdoba, R., Ferreiros, J.: A spanish speech to sign language translation system for assisting deaf-mute people. In: Proceedings of Interspeech. pp. 1399–1402 (2006)

13. Vossen, P. (ed.): EuroWordNet. A multilingual database with lexical semantic networks. Kluwer Academic Publishers (1998)
14. Young, S.: ATK: An Application Toolkit for HTK. Engineering Department, Cambridge University (2007)