Adapting a Speech into Sign Language Translation System to a new domain

V. López-Ludeña[1], R. San-Segundo[1], C. González-Morcillo[2], J.C. López[2], E. Ferreiro[3]

[1]Speech Technology Group. E.T.S.I. Telecomunicación. UPM.
[2]Grupo de Sistemas Inteligentes Aplicados, Dpto. de Tecnologías y Sistemas de Información. UCLM.
[3]Fundación para la Supresión de la Barreras de Comunicación. Fundación CNSE
veronicalopez@die.upm.es

Abstract

This paper presents a methodology for adapting an advanced communication system for deaf people in a new domain. This methodology is a user-centered design approach consisting of four main steps: requirement analysis, parallel corpus generation, technology adaptation to the new domain, and finally, system evaluation. In this paper, the new considered domain has been the dialogues in a hotel reception. With this methodology, it was possible to develop the system in a few months, obtaining very good performance: good speech recognition and translation rates (around 90%) with small processing times.

Index Terms: Development methodology, User-centered methodology, Speech into Spanish Sign Language translation, Advanced communication system for Deaf

1. Introduction

When developing human-computer interaction systems, it is very important to meet a set of requirements in order to guaranty their usability and user acceptation. In this process, a good methodology is very important for dealing with the main aspects to be considered. This fact is more relevant when involved users have any kind of disability. Based on the experience in previous projects, authors propose a specific methodology for developing an advanced communication system for deaf people focused on a specific domain. This advanced communication system permits real face to face interactions between hearing and deaf people, allowing a natural interaction between them. This system is able to translate spoken Spanish into LSE (Lengua de Signos Española – Spanish Sign Language) and generating speech from LSE.

In recent years, several groups have shown interest in spoken language translation into sign languages, developing several prototypes: example-based [1], rule-based [2], grammar-based [3], full sentence [4] or statistical ([5]; SiSi system1;[6]) approaches. For LSE, it is important to remark the author experience developing speech into LSE (Spanish Sign Language) translation systems in several domains ([2]; [7]; [8]).

In order to eliminate the communication barriers between deaf and hearing people, it is necessary not only to translate speech into sign language [2] but also to generate spoken language from sign language, allowing a fluent dialogue in both directions. A great deal of effort has been made in recognizing sign language and translating it into spoken language by using a language translator and a TTS converter. The main efforts have focused on recognizing signs from video processing. The systems developed so far are very person or environment dependent [1], or they focus on the recognition of isolated signs ([10]; [11]) which can often be characterized just by the direction of their movement. A system for automatic translation of static gestures of alphabets and signs in American Sign Language is presented by using Hough transform and neural networks trained to recognize signs in [12]. In the Computer Science department of the RWTH Aachen University, P. Dreuw is making a significant effort into recognizing continuous sign language from video processing [13].

In parallel, the Deaf community has found a new communication alternative based on SMS (Short Message Service) languages. Bearing in mind this scenario, the advanced communication system developed in this paper includes the LSESpeak system [14], a new application for helping Deaf people to generate spoken Spanish. This application includes two main utilities: the first one is a spoken Spanish generator from LSE and the second one is an SMS language to spoken Spanish translation system.

2. Methodology overview

The methodology presented in this paper is an adaptation of the Participatory Design methodology: one of the most used User-Centered Design approaches that follows the ISO standard Human-centered design for interactive systems: ISO 9241-210, 2010. Participatory design (known before as 'Cooperative Design') is a designing approach where all stakeholders (e.g. employees, partners, customers, citizens, and end-users) are involved actively in the design process. The main target is to guarantee that the final designed product meets their needs and it is usable.

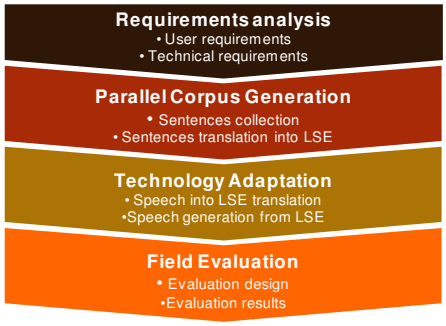

Figure 1: *Diagram of the methodology*

This methodology consists of the following phases or steps (Figure 1): requirements analysis, parallel corpus generation, technology adaptation and field evaluation. These steps will be

[1] http://www-03.ibm.com/press/us/en/pressrelease/22316.wss

explained in next sections. Section 3 describes the requirement analysis (user and technical requirements). Section 4 explains how the parallel corpus was generated. In section 5 the adaptation of the technology is described. And finally section 6 explains briefly the evaluation carried out.

3. Requirement analysis

For this step, it is necessary to define the domain where the advanced communication system will work. In this case, the new domain consists of the spoken dialogues between deaf customers and a receptionist in a hotel reception.

3.1. User requirements

According to the Survey of Disability, Personal Autonomy and Dependency Situations (EDAD, 2008) from INE (Spanish Statistic Institute), there are 1,064,100 deaf people in Spain. Deafness brings about significant communication problems: most deaf people have problems when expressing themselves in oral languages or under-standing written texts. Their communication barriers produce that 47% of deaf population has not studies or even is illiterate (INE -Spanish Statistic Institute- 1999 y MEC -Science and Education Ministry- 2000/2001). These aspects support the need to generate new technologies in order to develop automatic translation systems for helping in the hearing to deaf people communication.

In order to get the user requirements, two Participatory Design workshops where organized including deaf customers, hotel receptionists and researchers from all the project partners.

The first workshop was organized as a data collection and a brainstorming about the most frequent needs for deaf customers when they are in a hotel. As a result of this workshop, an initial report was generated including all the services offered by the hotel (hosting, parking, restaurant, internet, etc.), a typical structure and functioning of a hotel (check-in, check-out, scheduling, services, extra activities, accessibility, etc.) and specific needs for deaf people (visual alarms or visual alarm clock service, etc.).

The second workshop was focused on selecting the most important aspects in this domain (hotel reception). After this sorting, the most important (relevance for deaf users) services were selected for being addressed by the automatic system.

3.2. Technical requirements

An important challenge of the project is to achieve a minimum level of technical performance, because the acceptation depends strongly on this quality. Based on previous experiences ([2]; [7]), the technical researchers have defined the following technical requirements:

- The speech recognition system must provide a recognition rate higher than 90% in the selected application domain. If that rate is not reached with speaker-independent models, an adaptation process will be performed for each speaker involved in the evaluation in order to guarantee that rate.
- It is also necessary a translation rate higher than 90% for the specific domain targeted in the project. These performance constraints are necessary to guaranty a dynamic hearing-deaf dialogue (without many repetition turns).
- Finally, the avatar intelligibility must be higher than 90% when representing the signs: recognition rate from deaf people.

In order to guarantee these technical requirements, it will be required a Spanish-LSE parallel corpus with a significant number of sentences in the specific domain. Based on previous experiences, around 500 sentences are required containing less than 1,000 Spanish words and less than 200 signs in LSE.

4. Parallel corpus generation

This collection has been obtained with the collaboration of the Hotel "Intur Palacio de San Martín" of Madrid. Over several weeks, the most frequent explanations (from the receptionist) and the most frequent questions (from customers) were taken down. In this period, more than 1,000 sentences were noted and analyzed. Not all the sentences refer to the main services selected in the previous step, so sentences had to be selected manually. Finally, 500 sentences were collected: 276 pronounced by receptionists and 224 by customers. This corpus was increased to 1,677 by incorporating different variants for Spanish sentences (maintaining the meaning and the LSE translation). These sentences were translated into LSE, both in text (sequence of glosses) and in video, and compiled in an excel file.

The main features of the corpus are summarized in Table 1. These features are classified according to the speaker (the receptionist or the customer).

Receptionist	**Spanish**	**LSE**
Sentence pairs	937	
Different sentences	770	243
Running words	2394	3349
Vocabulary	772	389
Customer	**Spanish**	**LSE**
Sentence pairs	741	
Different sentences	594	200
Running words	4091	2394
Vocabulary	594	277

Table 1: *Main statistics of the parallel corpus*

5. Technology adaptation

The Advanced Communication System is composed on two main modules: a spoken Spanish into LSE (Lengua de Signos Española – Spanish Sign Language) translation module and a spoken Spanish generator from LSE.

5.1. Speech into LSE translation

This module is used to translate the receptionist utterances into LSE. Figure 2 shows the module diagram.

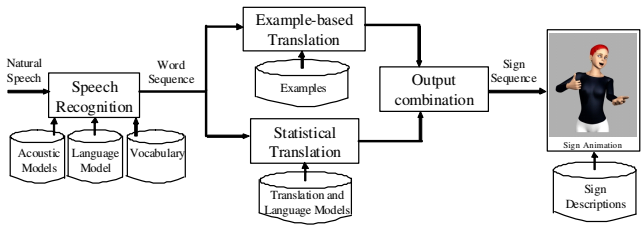

Figure 2: *Speech into LSE translation system*

5.1.1. *Speech Recognition*

The automatic speech recognizer (ASR), converts natural speech into a sequence of words (text). It is used an HMM (Hidden Markov Model)-based system able to recognize continuous speech. It has been developed at Speech Technology Group (GTH-UPM[1] [8]).

In order to adapt the speech recognizer to a new domain, the vocabulary and language models must be updated using the Spanish sentences collected for each application domain. The speech recognizer includes an acoustic adaptation module for adapting the acoustic models to a new specific environment, a new speaker, or a new Spanish accent. Also, when generating automatically the vocabulary and a language model for the speech recognizer, a new module has been included for introducing source language variants, increasing the speech recognizer flexibility.

5.1.2. *Language Translation*

The natural language translation module converts a word sequence into a sign sequence. It has a hierarchical structure divided into two main steps (Figure 3). In the first step, an example-based strategy is used to translate the word sequence in order to look for the best possible match. If the distance with the closest example is lower than a threshold (Distance Threshold), the translation output is the same as the example translation. But if the distance is higher, a background module based on a statistical strategy translates the word sequence [8].

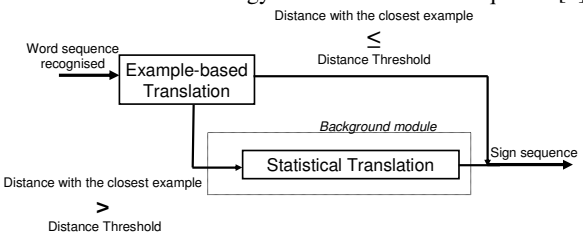

Figure 3: *Language translation module*

An example-based translation system uses a set of sentences in the source language (Spanish) and their corresponding translations in the target language (LSE). In order to determine whether one example is equivalent to the sentence to be translated, the system computes a heuristic distance between them. If the distance is lower than a threshold (25%), the translation output will be the same as the example translation. But if the distance is higher, it is necessary to have a background module. For adapting the example-based translation module, it is necessary to process the receptionist part of the parallel corpus for generating the translation examples in a new domain. In order to make the examples more effective, it is possible to generalize them [15]. The translation examples are generalized by replacing common words (like numbers, months, weekdays, etc.) with categories. This way, generalized examples can translate a higher number of sentences.

The background translation module consists of a statistical translator based on statistical phrase translation and language models [16]. When adapting this module to a new domain, it is necessary to adapt both translation and language models using the receptionist part of the parallel corpus. If the corpus is bigger, more reliable the model probabilities are.

5.1.3. *Sign Language Representation*

The Sign Language Representation module uses a declarative abstraction module used by all of the internal components. This module used a description based on XML, where each key pose configuration is stored defining its position, rotation, length and hierarchical structure. We have used an approximation of the standard defined by H-Anim (Humanoid Working Group ISO/IEC FCD 19774:200x). In terms of the bones hierarchy, each animation chain is composed by several « joint » objects that define transformations from the root of the hierarchy.

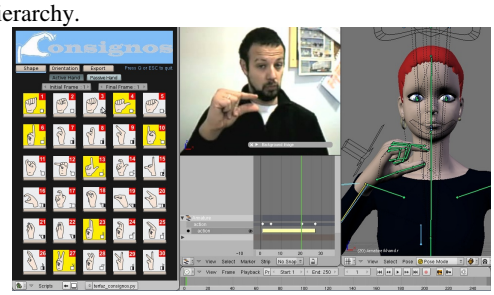

Figure 4: *Sign editor*

In order to increase its adaptability, the sign animation module (Figure 4) includes a new version of the sign editor based on the use of predefined static poses for hand shape and orientations. It incorporates predefined positions and orientations for reducing the sign specification time.

5.2. Speech Generation from LSE

The speech generator from LSE converts the deaf customer questions into spoken Spanish. We have integrated the LSESpeak system [14]. LSESpeak is made up of two main tools (Figure 5). The first one is a new version of an LSE into Spanish translation system [17], and the second one is an SMS to Spanish translation system, because Spanish deaf people become familiar with SMS language. Both tools are made up of three main modules.

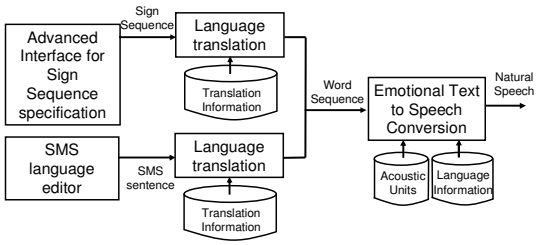

Figure 5: *LSESpeak system*

The first module is an advanced interface in which it is possible to specify an LSE sequence or an SMS message. The second module is a language translator for converting LSE or SMS into written Spanish. In both cases, the language translating modules are open-source phrase-based translation modules based on the software released at the 2011 EMNLP Workshop on Statistical Machine Translation[2]. In the case of the SMS into text translation module, it has been necessary to

[1] http://lorien.die.upm.es

[2] http://www.statmt.org/wmt11/

include pre-processing and post-processing modules in order to deal with some specific characteristics.

Finally, the third module is an emotional text to speech converter based on Hidden Semi-Markov Models (HSMMs) in which the user can choose the voice gender (female or male), the emotion type (happy, sad, angry, surprise, and fear) and the Emotional Strength (ES) (on a 0-100% scale).

6. Field evaluation

The final step of the methodology is the field evaluation. The communication system must be tested with real users in a real scenario. This step is divided in two main tasks: evaluation design and field evaluation with real users.

6.1. Evaluation Design

The evaluation design was performed in a workshop where end-users (deaf people and receptionists), researchers and developers work together to define the evaluation plan. This plan includes the main characteristics of the testing users, the main resources necessary for the evaluation, the consent form, the scenarios to simulate and the objective and subjective measurements.

6.2. Evaluation Results

The evaluation was carried out over one day. The speech recognizer was adapted to the receptionist involved in the evaluation. The system was evaluated by four deaf customers (two females and two males) who interact with one receptionist at the reception of the Intur Palacio San Martin Hotel. The deaf customer's ages ranged between 26 and 47 years. All the customers said that they use a computer every day or every week, and only half of them had a medium-high understanding level of written Spanish.

Before using the developed system, the deaf customers observed several signs represented with the avatar and they were asked to identify them considering this specific domain. After that, they were asked to interact with the receptionist using the advanced communication system in the scenarios designed in the evaluation plan preparation. After the interactions, the deaf customers were asked several specific questions about the information provided by the receptionist.

AGENT	MEASUREMENT	VALUE
System	Word Error Rate	6.7%
	Sign Error Rate (after translation)	10.7%
	Average Recognition Time	3.1 sec
	Average Translation Time	0.002 sec
	Average Signing Time	4.1 sec
	# of dialogues	11

Table 2: *Objective measurements of the evaluation*

Table 2 shows the objective measurements of the evaluation. The WER (Word Error Rate) for the speech recognizer is 6.7% being small enough to guarantee a low SER (Sign Error Rate) in the translation output: 10.7%. On the other hand, the time needed for translating speech into LSE (speech recognition + language translation + sign representation) is around 7 seconds allowing an agile dialogue. This performance fits the technical requirements defined.

The questionnaires summarize the recognition accuracy based on the number of attempts for isolated signs and for questions about the dialogues. For isolated signs, the recognition rate in the first attempt is very high (close to 90%, technical requirement defined) but for the dialogues, this percentage was worse, close to 60%. The main problems were related to the recognition of some signs: there were problems on the orientation of several signs and the discrepancy about what sign to choose for presenting one concept. LSE (Lengua de Signos Española) is a very young language (it is official since 2007) and there is a very high variability between different regions in Spain. These differences affect sign representation and the structure of the sign language sentences.

Human recognition rate depending on the number of attempts			
	1st	2nd	3rd
Isolated signs	87.5%	10.0%	2.5%
Questions	62.5%	25.0%	12.5%

Table 3: *Subjective measurements in the questionnaires*

Finally, some objective measurements of the spoken Spanish generation module are included in Table 4.

AGENT	MEASUREMENT	VALUE
System	Translation rate	98.0%
	Average translation time	0.001 sec
	Average time for text to speech conversion	2.1 sec
	Time for gloss sequence specification.	18.0 sec
	# of dialogues	11

Table 4: *Objective measurements for evaluating the Spanish generator from LSE*

As it is shown, the good translation rate and the short translation time make available the possibility to use this system in real conditions. The user needed less than 20 seconds to specify a gloss sequence using the interface. This time is not very high considering that the deaf customer had only few minutes to practice with the visual interface before the evaluation. With more time for practicing, this time would be reduced.

7. Conclusions

This paper has presented a developing methodology whose main advantage is that the users (hotel receptionist and deaf customers) are involved in the most important steps: requirement analysis, parallel corpus generation and field evaluation. Another important advantage is that the technology adaptation is almost automatic from the parallel corpus. With this methodology, it has been possible to develop the system in a few months, obtaining very good performance. On the other hand, the main disadvantage is that the methodology is sequential and the technology adaptation depends on the parallel corpus generation.

Acknowledgements

The work leading to these results has received funding from the European Union under grant agreement n° 287678. It has also been supported by SAILSE (IMSERSO), TIMPANO (TIN2011-28169-C05-03), ITALIHA (CAM-UPM), INAPRA (MICINN, DPI2010-21247-C02-02) and MA2VICMR (CAM, S2009/TIC-1542) projects.

References

[1] Morrissey S., and Way A.,. 2005. “An example-based approach to translating sign language”. In Workshop Example-Based Machine Translation (MT X–05), pages109–116, Phuket, Thailand, September.

[2] San-Segundo R., Barra R., Córdoba R., D’Haro L.F., Fernández F., Ferreiros J., Lucas J.M., Macías-Guarasa J., Montero J.M., Pardo J.M, 2008. “Speech to Sign Language translation system for Spanish”. Speech Communication, Vol 50. 1009-1020. 2008.

[3] Marshall, I., Sáfár, E. (2005) “Grammar Development for Sign Language Avatar-Based Synthesis”, In Proceedings HCII 2005, 11th International Conference on Human Computer Interaction (CD-ROM), Las Vegas, USA, July 2005.

[4] Cox, S.J., Lincoln M., Tryggvason J., Nakisa M., Wells M., Mand Tutt, and Abbott, S., 2002 “TESSA, a system to aid communication with deaf people”. In ASSETS 2002, pages 205-212, Edinburgh, Scotland, 2002.

[5] Bungeroth J., Ney, H., “Statistical Sign Language Translation. In Workshop on Representation and Processing of Sign Languages”, LREC 2004, 105-108.

[6] Morrissey S., Way A., Stein D., Bungeroth J., and Ney H., 2007 “Towards a Hybrid Data-Driven MT System for Sign Languages. Machine Translation Summit (MT Summit)”, pages 329-335, Copenhagen, Denmark, September 2007.

[7] San-Segundo, R., Montero, J.M., Córdoba, R., Sama, V., Fernández, F., D’Haro, L.F., López-Ludeña, V., Sánchez D., and García, A., 2011. Design, development and field evaluation of a Spanish into sign language translation system. Pattern Analysis and Applications. Pattern Analysis and Applications. Volume 15, Issue 2, pp 203-224.

[8] López-Ludeña, V., San-Segundo, R., González Morcillo, C., López, J.C., Pardo, J.M.,. 2013b. “Increasing adaptability of a speech into sign language translation system”. Expert Systems With Applications. Volume 40, Issue 4, March 2013, Pages 1312–1322.

[9] Vogler, C., Metaxas, D. 2001: “A Framework for Recognizing the Simultaneous Aspects of ASL”. CVIU 81(3), 358–384 (2001).

[10] Wang, S.B., Quattoni, A., Morency, L.-P., Demirdjian, D., Darrell, T.: Hidden Conditional Random Fields for Gesture Recognition. In Proc. of CVPR, June 2006, vol. 2, pp. 1521–1527 (2006).

[11] von Agris, U., Schneider, D., Zieren, J., Kraiss, K.-F.: 2006. “Rapid Signer Adaptation for Isolated Sign Language Recognition”. In Proc. of CVPR Workshop V4HCI, New York, USA, June 2006, p. 159 (2006).

[12] Munib, Q., Habeeb, M., Takruri, B., Al-Malik, H. 2007 “American sign language (ASL) recognition based on Hough transform and neural networks”. Expert Systems with Applications, Volume 32, Issue 1, January 2007, Pages 24-37.

[13] Dreuw, P., Stein D., and Ney H. 2009. “Enhancing a Sign Language Translation System with Vision-Based Features”. In Proc. of Special Issue Gesture Workshop 2007, LNAI, number 5085, pages 108-113, Lisbon, Portugal, January 2009.

[14] López-Ludeña, V., Barra-Chicote, R., Lutfi, S., Montero, J.M., San-Segundo, R., 2013a “LSESpeak: A spoken language generator for Deaf people”. Expert Systems With Applications. Volume 40, Issue 4, March 2013, Pages 1283–1295

[15] Brown. RD., 2000. "Automated Generalization of Translation Examples". In Proceedings of the Eighteenth International Conference on Computational Linguistics (COLING-2000), p. 125-131. Saarbrücken, Germany, August 2000.

[16] López-Ludeña, V., San-Segundo, R., Montero, J.M., Córdoba, R., Ferreiros, J., Pardo, J.M., 2012 "Automatic Categorization for Improving Spanish into Spanish Sign Language Machine Translation" Computer Speech and Language, Volume 26, Issue 3, June 2012, Pages 149–167.

[17] San-Segundo, R., Pardo, J.M., Ferreiros, F., Sama, V., Barra-Chicote, R., Lucas, JM., Sánchez, D., García. A., 2010. “Spoken Spanish Generation from Sign Language” Interacting with Computers, Vol. 22, No 2, pp. 123-139, 2010.