3D Modeling of the Mexican Sign Language for a Speech-to-Sign Language System

Santiago-Omar Caballero-Morales and Felipe Trujillo-Romero

Postgraduate Division, Technological University of the Mixteca, Oaxaca, Mexico

{scaballero, ftrujillo}@mixteco.utm.mx

Abstract. There are many people with communication impairments, deafness being one of the most common of them. Deaf people use Sign Language (SL) to communicate, and translation systems (Speech/Text-to-SL) have been developed to assist such communication. However, since SLs are dependent of countries and cultures, there are differences between grammars, vocabularies, and signs, even if these come from places with similar spoken languages. In Mexico, work in this field is very limited, so any development must consider the characteristics of the Mexican-Sign-Language (MSL). In this paper, we present a new approach to creating a Mexican Speech-to-SL system, integrating 3D modeling of the MSL with a multi-user Automatic Speech Recognizer (ASR) with dynamic adaptation. The 3D models (avatar) were developed by means of motion capture of a MSL performer. Kinect was used as a 3D sensor for the motion capture process, and DAZ Studio 4 was used for its animation. The multi-user ASR was developed using the HTK and Matlab as the programming platform for a Graphical User Interface (GUI). Experiments with a vocabulary set of 199 words were performed to validate the system. An accuracy of 96.2% was achieved for the ASR and interpretation into MSL of 70 words and 20 spoken sentences. The 3D avatar presented clearer realizations than those of standard video recordings of a human MSL performer.

Keywords. Mexican sign language, automatic speech recognition, human-computer interaction.

Modelado 3D del lenguaje de señas mexicano para un sistema de voz-a-lenguaje de señas

Resumen. Hay muchas personas con problemas para comunicarse, siendo la sordera una de las más comunes. Personas con este problema hacen uso de Lenguaje de Señas (LSs) para comunicarse, y sistemas de traducción (Voz/Texto-a-LS) se han desarrollado para asistir a esta tarea. Sin embargo, porque los LSs son dependientes de países y culturas, hay diferencias entre gramáticas, vocabularios y señas, incluso si estos provienen de lugares con lenguajes hablados similares. En México, el trabajo es muy limitado en este campo, y cualquier desarrollo debe considerar las características del Lenguaje de Señas Mexicano (LSM). En este artículo, presentamos nuestro enfoque para un sistema de Voz-a-LS Mexicano, integrando el modelado 3D del LSM con un Reconocedor Automático de Voz (RAV) multi-usuario con adaptación dinámica. Los modelos 3D (avatar) fueron desarrollados por medio de captura de movimiento de un signante del LSM. Kinect fue usado como un sensor 3D para el proceso de captura de movimiento, y DAZ Studio 4 fue usado para su animación. El RAV multi-usuario fue desarrollado usando HTK y Matlab fue la plataforma de programación para la Interfaz Gráfica de Usuario (GUI). Experimentos con un vocabulario de 199 palabras fueron realizados para validar el sistema. Una precisión del 96.20% fue obtenida para el RAV e interpretación en vocabulario del LSM de 70 palabras y 20 frases habladas. Las realizaciones del avatar 3D fueron más claras que aquellas de grabaciones de video de un signante humano del LSM.

Palabras clave. Lenguaje de señas mexicano, reconocimiento automático de voz, interacción humano-computadora.

1 Introduction

Research on spoken language technology has led to the development of Automatic Speech Recognition (ASR) systems, Text-to-Speech (TTS) synthesis, and dialogue systems for interaction with artificial agents (e.g., robots). These systems are now used for different applications such as in mobile devices for voice

ISSN 1405-5546
DOI: 10.13053/CyS-17-4-2013-011

dialing and web navigation, information retrieval, dictation [1, 2], translation [3], and assistance to handicapped people [4].

Recently, ASR technology has been used for language learning. Examples of this application can be found in [5] for English, [6] for Spanish and French among others, and in [7] for Sign Languages (SLs). In general, these systems enable the user to communicate in different languages, and communication abilities are essential for people's lives.

Sign Languages (SLs) are languages that use a system of manual, facial, and other body movements as the means of communication, especially among deaf people. Thus, SLs are languages that, instead of using sounds (articulation of phonemes) to create words, use gestures. Important elements of a SL are hand configurations, facial and corporal expressions, and the Sign Writing (SW), which is the way used to write a SL. Hand configurations are positions made with the hands to create signs. Some words are made with only one hand, while others are made with both hands. Hand alphabets, in which a hand configuration represents a letter in the alphabet, are used to "spell" words of spoken languages with no representation in the SL. SLs are not universal, and so, there is no single SL for all the users around the world. These are independent languages, each with its own grammar, syntax, morphology, and semantics.

As people in general require time to learn a language, deaf people require time to learn a SL, which is important to communicate with other deaf or hearing people. Commonly, professional signants are required to assist communication and learning among deaf and hearing people. A technological tool that could contribute to these tasks in real time would be of valuable support for the community.

Research has been done to provide such tools, and many works with special features have been produced. San-Segundo *et al*. [8] developed a Spanish Speech-to-SL translation system to assist deaf people. This system achieved an error rate of 27.3%, and integrated an Automatic Speech Recognition (ASR) system with a 3D avatar to show the realization of the recognized speech into Spanish Sign Language (SSL). Another approach was shown in [9] by Baldassarri *et al*. Their Speech-to-SL translation system was based on the grammar rules of the Spanish language and considered the morphological and syntactical relationships of words in addition to their semantics. Another work of interest is the one developed by López-Colino and Colás [10]. They established a first approximation to the automatic synthesis of Spanish speech into SSL by construction of classifiers. These generated sequences of SSL movements were represented by means of a 3D avatar. Massó and Badia [11] used a morpho-syntactic approach to generate a statistical translation machine. Although in [11] the Catalán language was used instead of Spanish, the results can be useful for the development of a Spanish system due to similarities between these languages. Also, since both languages are spoken in the same region of Spain, many signs are very similar.

A common approach in the works mentioned previously is the development of a Speech-to-SL translator, which interprets sounds into hand configurations of a SL. Such system can assist real-time communication between deaf and hearing people as well as contribute to learning activities. For the Mexican community, we consider that developing such a system can be a valuable contribution because few works have addressed this issue locally, the most significant work being the Dictionary of the Mexican Sign Language (Diccionario Español - Lengua de Señas Mexicana, DIELSEME) [12]. Note that the usability of this resource for real time communication is limited due to restrictions typical for any dictionary.

We present the development of a Mexican Speech-to-SL translator system considering some particular issues related to its design. For example, all SLs are different even if they come from places with similar spoken languages. Thus, a given Spanish word may be represented by different signs in Mexico, Colombia or Spain, even though Spanish is the language of these countries and the word is written exactly in the same way. Reusing any SL library of previous works for the Spanish language would be very inappropriate as Mexico has its own SL, the Mexican Sign Language (MSL). Thus, in order to develop a system as those developed in the

reviewed works, such important resources as a 3D avatar or models of the MSL must be created. 3D modeling of the MSL is a significant extension of works related to only modeling hand gestures with limited interpretation as in [13] and [14]. Hence, the information presented in this paper can also contribute to the development of recognition systems for the MSL or other SLs.

The development of the system is presented as follows. In Section 2 we provide the details of the Kinect-based motion capture procedure performed for the 3D modeling of the MSL and avatar creation. In Section 3, the details of the multi-user ASR system for the Mexican Spanish language are shown. In Section 4, the details of the architecture of the Speech-to-SL translator and the interface for the system are presented. Finally, the performance results of the integrated interface are shown in Section 5, and in Section 6 conclusions and future plans for this project are discussed.

2 3D MSL Modeling and Avatar Creation

As mentioned, an important part of the Speech-to-SL translator is the set of hand configurations which define each MSL unit in the vocabulary set of the system. Some recent works are based on video libraries, and sets of static captures of signs as presented in [13] and [14]. However, this restricts the modeling and its application for recognition purposes with a bigger set of symbols. For example, in [13] only five different signs were captured, while in [14] only six were modeled. Especially for SLs, particular attention must be given to fine movements of hands and fingers.

Recent development of entertainment devices such as Kinect and Wii Remote for motion sensing are an alternative for applications in SL processing. 3D motion capture was considered as it enables information extraction of all the movements involved in the sign performing process. Also, integrating this information into a 3D avatar was found to be more suitable for understanding by deaf users [7, 15].

For this work, 3D capture of MSL symbols was performed using the Microsoft Kinect [16] as a sensor. Kinect makes the potential use of motion capture more viable given its low price in the market. However, its resolution permits to capture body motion only, as hands and face gestures require more accurate sensing hardware. However, for the initial modeling and creation of a 3D avatar, it provided significant information. In Fig. 1 the general steps of this process are presented. Each step is described in the following sections.

2.1 3D Data Acquisition

The Kinect sensor is a device that belongs to range-finding technologies which use time-of-flight to determine the distance to objects present in a scene. This kind of sensor captures three-dimensional information by measuring the time it takes by an infrared (IR) signal to travel to an object and be reflected back to the signal's source.

A laser source emits invisible light which passes through a filter and is scattered into a pattern of small points which is projected onto the environment in front of the sensor. The reflected pattern is then detected by an IR camera and analyzed. From the emitted pattern, lens distortion, and the distance between the emitter and the receiver, the distance to each point can

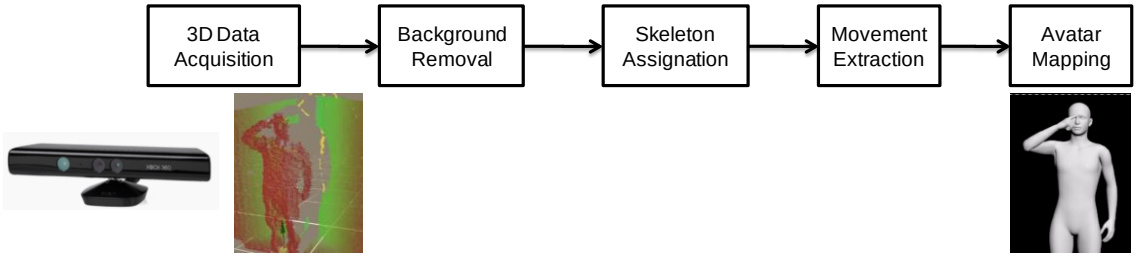


Fig. 1. Development steps of the 3D avatar

(a) (b)

Fig. 2. (a) Depth image from Kinect sensor, (b) 3D points from (a)

be estimated. This process is done internally in the Kinect and the final depth image is obtained in a direct way.

An example of a depth image of the performance of a MSL signant is presented in Fig. 2(a). The image in Fig. 2(b) shows the cloud of 3D points obtained by means of a transformation between the depth image and the Kinect's parameters of calibration.

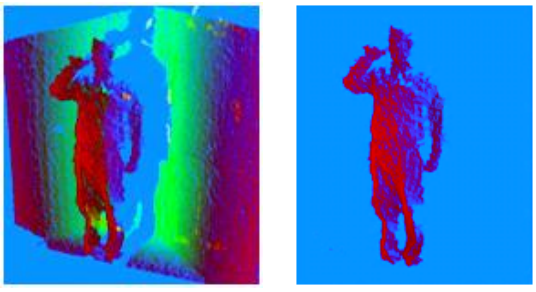
2.2 Background Elimination

Due to the fact that at the moment of motion capture the information of the background is also considered, this information interferes with the data of the MSL realizations. Hence, it is necessary to eliminate the three-dimensional data related to the background of the scene.

Fig. 3(a) shows the raw 3D data from the Kinect sensor where the foreground and the background are present, see Fig. 2(b). There are two main procedures used in order to obtain the data of interest: (1) subtraction, and (2) threshold.

In the first case, an image of reference (background) is taken before acquisition of data. Then, when the MSL performance is captured, the reference data is subtracted, eliminating the information associated to the background, thus keeping the MSL data only.

In the second case, a threshold is defined with the purpose of eliminating all the information beyond a given point or distance. Thus, for the MSL performance of interest, only the data associated to the foreground will be captured. We used the first method because (a) all the images were taken using the same background, and (b) its implementation is easier as the subject can be posed anywhere within the range of the Kinect sensor. The result is presented in Fig. 3(b).



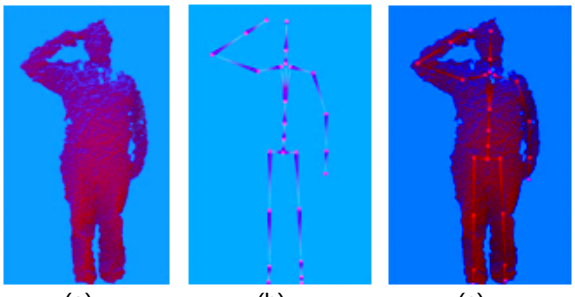
(a) (b)

Fig. 3. (a) Original capture, (b) capture with elimination of background

2.3 Skeleton Assignation

Once the background is eliminated (or removed), a virtual skeleton is assigned to the clean data. The assignation of the skeleton is required to perform tracking of the signant's movements and to map them to a 3D avatar (MSL models).

For this process it is assumed that only the information related to the MSL signant is available and the whole body is perceived, see Fig. 4(a). In order to assign the skeleton to the 3D points that represent the signant, the Skeletal Viewer Walkthrough [16] of the Microsoft Kinect SDK was used. The information of the skeleton presented in Fig. 4(b) was then assigned to the cloud of 3D points of Fig. 4(a), resulting in the composite data presented in Fig. 4(c), where it is possible to observe the 3D data with the skeleton.



(a) (b) (c)

Fig. 4. (a) Original 3D data, (b) skeleton, (c) skeleton integrated with the 3D data

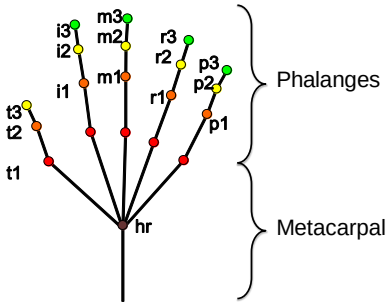


Fig. 5. Structure of the hand skeleton

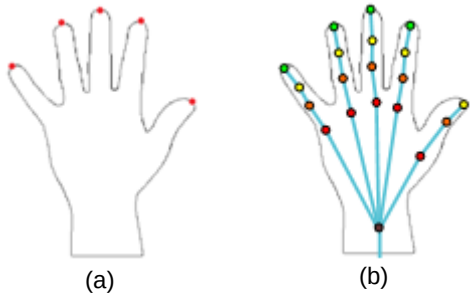


Fig. 6. Steps for tracking: (a) fingertip detection, and (b) assignation of hand structure

2.4 Finger Movement Extraction

Since the SDK of the Kinect sensor does not consider hand poses, it was necessary to develop a strategy to map the hand configurations and the finger movements of the MSL. This is not an easy task due to null information about the finger localization and movement obtained by using the Kinect sensor. Thus, we approached the modeling of the finger movements by means of prediction. A particle filter was used in order to estimate the localization and configuration of the fingers from an initial position. For this, a structure was developed to assign and manipulate both hand and fingers. This structure allowed us to make the finger assignment to the 3D data which was acquired with the Kinect sensor.

In Fig. 5 we present the Hand Skeleton structure, where every point identifies a finger articulation. Using this structure, we generated a model to recover the different hand configurations of the MSL alphabet. We took the correspondences between the proportions of the finger's bones in the human hand [17] in order to get a good match between the 3D cloud of points and the hand skeleton.

With the assignation of this structure we partially resolved the finger's tracking problem. However, we needed a technique to make permanent tracking possible as the meaning of some MSL signs depends on the sequence of finger movements. This was also important due to the occlusions between the fingers and the position and orientation of the hand.

As previously mentioned, we used a particle filter to perform prediction as a technique for finger tracking like in [18] and [19]. The particle filter detects the fingertips and predicts the next position of the fingers even if there are occlusions.

The detection of fingertips is performed by computing the *k*-curvature of the fingers [20] and using the disparity image. This process is represented in Fig. 6(a) which shows the detection of fingertips, and Fig. 6(b) which presents the assignation of the hand structure.

2.5 Avatar Mapping

Finally, the assignation of the skeleton was used to perform tracking of the user's movements and to map them to a 3D avatar. In order to perform this, all the data from the skeletons related to positions was recorded, and files of type *bvh* were generated. The file format *bvh* was developed by Biovision, a motion capture services company, to provide motion capture data to their customers. The acronym *bvh* stands for *Biovision hierarchical data*.

These *bvh* files were then used with DAZ Studio [21] to map the MSL movements to a 3D model, and *Genesis* was the one used in this work. It is a simple model without any texture or gender. Additional animation was added to refine finger movements required for some MSL signs.

Word models were created for the vocabulary presented in Table 3 and taken from the Spanish Dictionary of Mexican Sign Language (Diccionario Español - Lengua de Señas Mexicana, DIELSEME) [12].

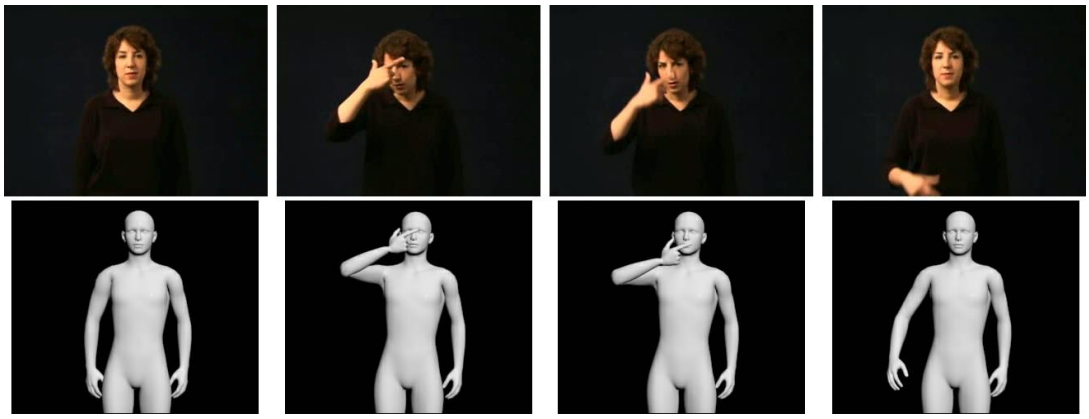


Fig. 7. Comparison of the DIELSEME video library and the 3D word-based avatar realizations of the word NOCHE (night)

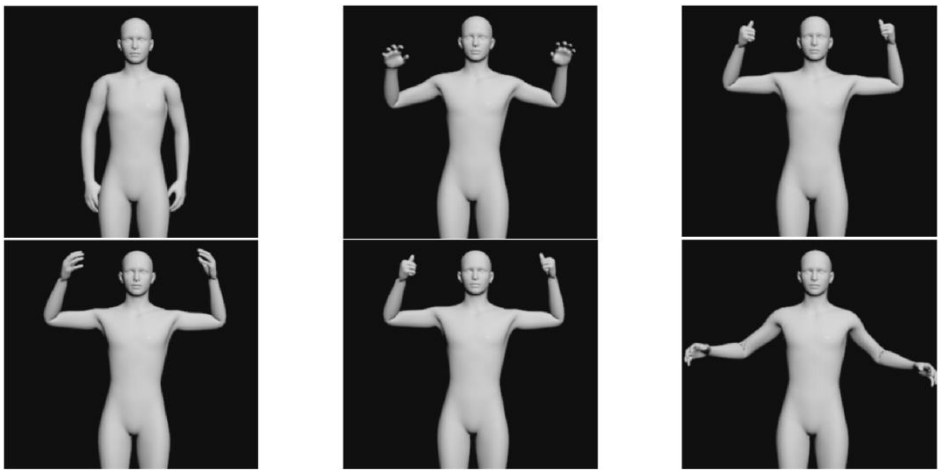


Fig. 8. Word-based MSL avatar model for the word *ADIOS* (goodbye)

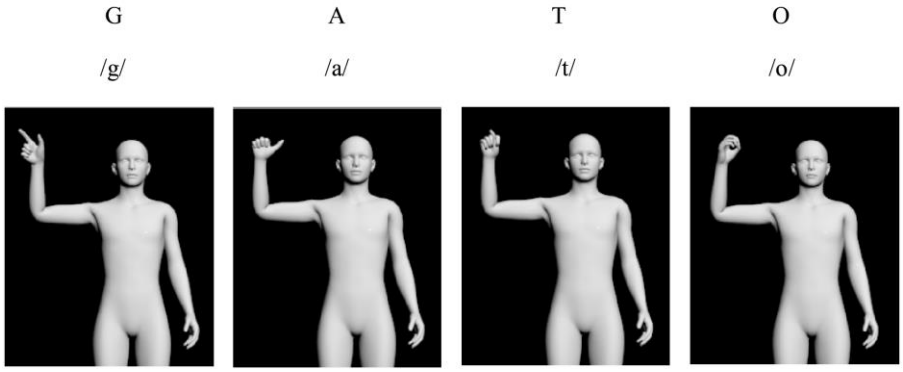


Fig. 9. Character-based MSL avatar model for *GATO* (cat)

ISSN 1405-5546
DOI: 10.13053/CyS-17-4-2013-011

In Fig. 7 some examples of the 3D avatar are shown as well as its comparison to the DIELSEME library. As it can be observed, the avatar has higher definition in the signs that require fine finger movements. The final animations for the avatar were satisfactory as evaluated by the MSL signant.

In order to generate an animated sequence for words not present in the selected vocabulary, the set of signs that represent each character of the Mexican alphabet (e.g., *A*, *B*, *C*, …, *Z*) was modeled. Thus, for the Speech-to-Sign language interaction system, word-based and character-based (for spelling) MSL avatar models were built. An example of a word-based model is the one presented in Fig. 7 (for the word *NOCHE*) and Fig. 8 (for the word *ADIOS*), where specific sequences of hand configurations are associated to each word. In contrast, in Fig. 9 the sequence of hand configurations associated to the characters that form the non-present (in the selected vocabulary) word *GATO* (cat) is given. Note that the sequence of these characters is not equal to the sequence defined for the whole word. This sequence is obtained from the phoneme transcription of the word obtained by the ASR system presented in the next section.

3 Multi-User Mexican ASR System

After the 3D avatars for the MSL vocabulary were built, we proceeded to develop an accurate Automatic Speech Recognition (ASR) system. This is very important for our proposal, as an appropriate sign translation relies on correctly recognized speech commands.

In order for a robust and accurate ASR system to be used by a variety of speakers, it must be trained with the speech samples from many speakers. Such system defined as Speaker Independent (SI) [23, 24] can be adapted to a specific user by means of speaker adaptation techniques such as MAP or MLLR [23, 24, 25]. However, for a Mexican ASR system, there are few resources (e.g., Speech Corpora) to build the components of a SI system. Thus, to develop the ASR module for the speech-to-sign system with limited resources, it was assumed that the robustness of a SI ASR system could be achieved with few training speakers if

Table 1. Background of training speakers for the ASR module

Age	Gender	Origin	Occupation
17	Male	Mexico City	Student
40	Male	Mexico City	Teacher
27	Male	Puebla	Technician
34	Male	Puebla	Teacher
55	Male	Oaxaca	Teacher
37	Female	Mexico City	Linguist
15	Female	Puebla	Student
50	Female	Puebla	Accountant
24	Female	Puebla	Student
39	Female	Oaxaca	Technician

- the training speakers were representative of the main speech features (tones, pronunciations, etc.) in a language;
- there were enough speech samples for acoustic modeling;
- the vocabulary was not large (< 1000 words);
- the effect of statistical *a-priori* information (such as that of the Language Model) were adjusted for decoding (recognition);
- speaker adaptation were performed “dynamically” while using the system.

The ASR module was aimed to achieve recognition accuracies over 95% for test vocabularies. In the following sections the details of the construction of the ASR components are presented.

3.1 Speech Training Corpus

A corpus was built to obtain different pronunciations of the phonemes in the Mexican Spanish language. A representative text for the training corpus was obtained from [22] which consisted of

- 49 words used by a speech therapist to assess intelligibility;
- a fragment of a narrative that consisted of 102 words;

– 16 especially designed sentences which were phonetically balanced. For new users, this text was used as stimuli to obtain speech data for “static” adaptation.

In total, the representative text for the speech corpus consisted of 205 different words. Because most of commercial ASR systems are built with acoustic models at the phonetic level, the representative text was transcribed at the phonetic level to perform modeling of phonemes. For this purpose, the phoneme alphabet defined by the Master in Hispanic Linguistics Javier Octavio Cuétara [26] was used. This is an extension of the well-known Mexbet alphabet for the Mexican Spanish language. The phoneme alphabet, together with the frequency of occurrences of each phoneme in the representative text, is shown in Fig. 10. As presented, the alphabet consisted of 27 phonemes, plus */sil/* used to model *silence*, and */sp/,* to model *short-pauses*.

The phoneme transcriptions that define each word in the representative text were automatically obtained with the tool TranscribeMex [27, 28]. This library was developed to phonetically label the Mexican Spanish Corpus DIMEx100 and uses the alphabet described in [26]. Labeling of speech data at the phonetic level was performed after it was recorded from a selected group of speakers. In contrast to [29], where only six speakers (3 male, 3 female) were considered, and [22] where only one speaker (male) was considered, the corpus was built with the speech from ten speakers (5 male, 5 female). Their details are shown in Table 1.

These speakers were recruited based on their accent and place of origin because the phoneme definitions of TranscribeMex correspond to the Mexican Spanish of the center region of Mexico [27, 26]. With the exception of two speakers, all speakers currently live or work in Mexico City or Puebla (which is a 1.5 hour car travel from Mexico City). The speakers from Oaxaca located in the southern region of Mexico (a 6.7-7.5 hour car travel from Mexico City), had no particular differences in their accents as they had lived in Mexico City sufficient time.

Each speaker read the representative text (stimuli) as follows: five repetitions of the 49-word list, three repetitions of the narrative, and one repetition of 16 balanced sentences. Speech was recorded with a Sony lcd-bx800 recorder with a sampling frequency of 44 kHz monaural in WAV format. This data was then labeled manually at the word (orthographic) and phonetic level with the tool *WaveSurfer* [30].

3.2 Functional Elements

The functional elements of the ASR system were implemented with HTK [24]. This software uses Hidden Markov Models (HMMs) [31] for acoustic modeling of speech. In Fig. 11 the specific HTK modules used for each element of the ASR system are shown.

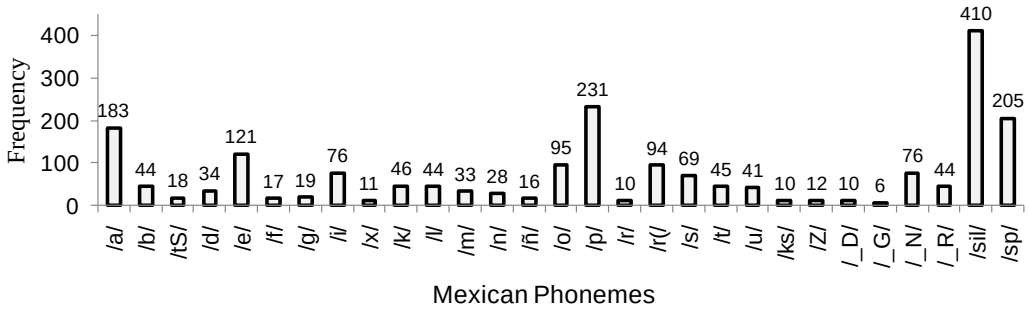


Fig. 10. Frequency of Mexican phonemes in the representative text [22]

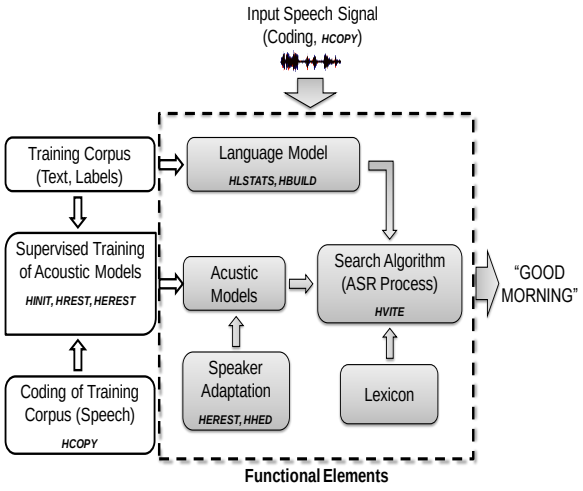


Fig. 11. Functional elements of the multi-user ASR system

The acoustic models consisted of HMMs with a standard three-state left-to-right architecture with eight mixture Gaussian components per state [23, 24, 31]. Since modeling was performed at the phonetic level, an HMM was built for each of the Mexican phonemes shown in Fig. 10. Then, for supervised training of the HMMs, the Speech Training Corpus was coded into Mel Frequency Cepstral Coefficients (MFCC's). The front-end used 12 MFCC's plus energy, delta, and acceleration coefficients [24].

The Language Model (LM) represents a set of rules or probabilities that restricts the sequence of words decoded by the ASR system to form valid phrases. The LM for this project consisted of bigrams (*N*-grams, where *N*=2) [23, 24], which were estimated from the orthographic transcriptions of the Training Speech Corpus.

The Lexicon, or dictionary, was built with TranscribeMex while we were labeling the Training Speech Corpus. This component only specifies the sequence of phonemes that define a word. For the Speech-to-MSL translator, this component provides the sequences of hand alphabets required to "spell" a non-present word in the MSL word-based database. However, this is not a straightforward process. It is important to mention that an alphabet letter is not equal to a phoneme. This is because a phoneme represents a sound or pronunciation, which is different from the text representation of a word or character. For example, in the words *PERRO* (dog) and *ROTO* (broken) the alphabet letters *RR* and *R* are pronounced as represented by the phoneme */r/* (*R* with strong pronunciation). Note that *RR* and *R* are placed before a vowel. For the words *FUERTE* (strong) and *SUERTE* (luck), the *R* has a strong pronunciation but it is shorter than in the previous words, and besides, it is placed after a vowel. The variation of *R* for this case is represented by the phoneme */_R/*, observing that the alphabet letter is still *R*. For the words *GRACIAS* (thanks) and *FRIO* (cold), *R* has a softer pronunciation which is represented by the phoneme */r(/*. On the other hand, *H* has no phoneme representation in Mexican Spanish.

Because of this situation, the dictionary built with TranscribeMex was revised to address the issues of phoneme and text representation for the Mexican Spanish alphabet for the non-present words. Nevertheless, for this work most of the alphabet letters could be associated to a single phoneme, leading to a straightforward equivalence as in the case of vowels. However, this should not be considered a rule for other languages.

The integration of all the elements to perform the ASR process (finding the sequence of words that best match the acoustic signal) is realized with the Viterbi algorithm [23].

To accomplish a multi-user ASR, Maximum Likelihood Linear Regression (MLLR) [25] was used as a speaker adaptation technique. For this task, 16 balanced sentences described in [22] were used as stimuli. A regression class tree with 32 terminal nodes was used for the HTK implementation of MLLR adaptation [24, 25].

3.3 Recognition Performance

A set of 20 spontaneous sentences of 2-7 words (mean of 4.1 words per sentence) were used as stimuli for testing the multi-user ASR system. These were used also as a source text for the Lexicon and LM components. 10 randomly selected people were asked to participate as test subjects and each one read 16 adaptation sentences prior to use the ASR system.

The measure of performance was the Word Recognition Accuracy (WAcc) [24] which is

defined as *WAcc = (N-D-S-I)/N* where *N* is the total number of elements (words) in the stimuli text; *D* and *I* are the number of elements deleted and inserted in the ASR's word output; and *S* the number of elements from the stimuli substituted by a different word in the output word sequence.

The results of the ASR performance are presented in Table 2. An overall recognition accuracy of 98.41% was achieved for the test speakers (98.66% correctly recognized words). In view of these results it was considered that the performance was satisfactory for the purpose of the project.

4 Mexican Speech-to-SL Translator

The Speech-to-SL translator links the recognized spoken words to specific sequences of MSL models. The structure of this system is shown in Fig. 12.

The translator searches in a MSL Database the 3D models that correctly match the recognized spoken word(s). If the recognized word(s) is (are) found in the MSL Database, then the translator proceeds to display the sequence of associated MSL models for that (those) word(s). Otherwise, if the word is not found in the MSL Database, then the word is spelled, and the sequence of models consists of the MSL models associated to each letter of the word(s).

For this work, the word-based MSL database consists of the vocabulary words presented in Table 3 (70 words in total). For the character-based MSL database, a 3D model was built for each of the alphabet letters in the Mexican Spanish language. In Table 4 the letters of the considered alphabet are presented, which led to the creation of 30 character-based MSL 3D models.

Table 2. Performance of the ASR system when tested by 10 randomly selected users

Age	Gender	Origin	Occupation	N	D	S	I	% WAcc
23	Male	Oaxaca	Student	82	0	2	0	97.56
36	Female	Mexico City	Teacher	82	1	0	1	97.56
28	Male	Oaxaca	Student	82	0	0	0	100.00
32	Male	Puebla	Teacher	82	1	1	0	97.56
27	Male	Oaxaca	Student	82	0	0	0	100.00
21	Male	Mexico City	Student	82	0	0	0	100.00
21	Female	Puebla	Student	82	1	2	0	96.34
20	Male	Oaxaca	Student	82	0	0	0	100.00
35	Female	Mexico City	Teacher	82	1	0	1	97.56
19	Male	Oaxaca	Student	82	1	1	0	97.56
				820	5	6	2	98.41

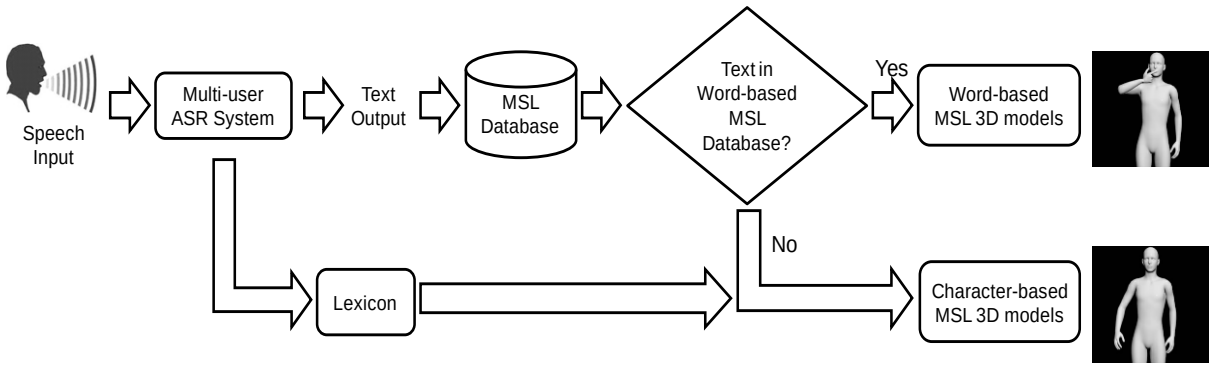


Fig. 12. Structure of the Mexican Speech-to-SL translator

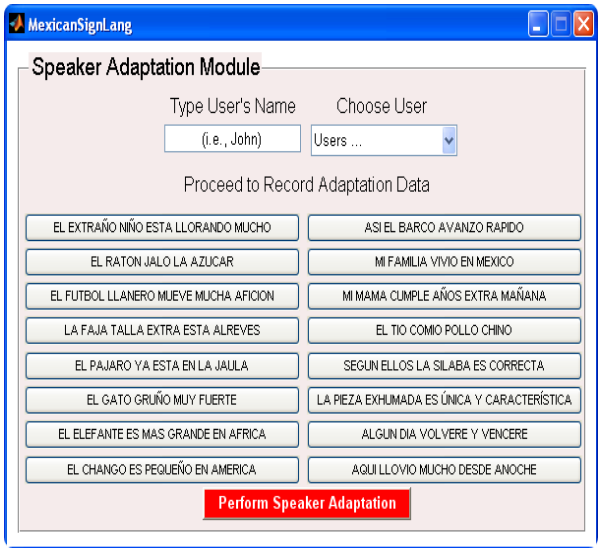


Fig. 13. Interface for user registration and speaker adaptation

4.1 Graphical User Interface

The ASR system, the translator, and the 3D avatars were integrated within a Graphical User Interface (GUI). The latter was developed using the GUIDE toolbox of Matlab (version R2008a), and was designed to dynamically add users and vocabulary to the translator system. In Fig. 13 the first module of the GUI, which performs the *user registration* and the "static" *speaker adaptation* tasks, is shown. This is an adaptation of the interface presented in [22].

Initially, a new user must type her name in the field *Type User's Name*. When the user presses "Enter" to add the name, the system automatically creates the user's directories to perform adaptation and updates her in the system's database. To start the adaptation task, the speaker is selected from the registered user's list in *Choose User*.

Then, by pressing each of the 16 buttons which are labeled with the stimuli text, the user can record the corresponding speech. These buttons turn "red" when pressed, which indicates that recording is being performed, and turn back to "white" when pressed again to stop the recording process. Internally, the system has the phonetic transcriptions of the stimuli text. By pressing *Perform Speaker Adaptation*, the system manages the HTK library to use the transcriptions and the recorded speech data to perform MLLR adaptation for the selected speaker.

In Fig. 14 the GUI of the Speech-to-SL translator is shown. Initially, the user selects her name from the list in the field *Choose User*. The system then automatically loads the adapted acoustic models for that user. The button *Update Vocabulary* builds the LM and Lexicon for the system with the vocabulary words shown in *Vocabulary*.

Additional words can be added to this list by typing them in the field *Add New Vocabulary*. These must be in uppercase format. If the user wants to use this text as stimuli for speaker adaptation, she can press the button *Record for Adaptation*. By doing this, the system starts recording the user's speech (the button will turn "red" as in the case of the adaptation task of Fig. 13). Internally, the system converts each of these words into phoneme sequences that are stored in the personal registers for that speaker (together with her acoustic models and MLLR data). The speech samples are also stored.

This process can be performed as many times as required. When the user finishes, she just needs to press the button *Adapt* to start the internal process of re-adaptation: the interface updates the list of adaptation speech samples for that speaker, updates the phonetic labels, and re-adapts the user's HMMs with the additional speech data. This process is considered to be "dynamic" as the user continuously can add more adaptation data without any restriction [22].

Speech-to-SL translation starts with the task of speech recognition, which is enabled by pressing *Execute Speech Recognition*. The recognized words are displayed under this button. An important parameter for the ASR is the *grammar scale factor* (*G Factor*), which controls the influence of the LM over the recognition process. Usually, a G Factor of "5" is used [24], however this can be adjusted to achieve a required level of performance. For this work, a G Factor of "10" was used.

Fig. 14. Interface for the Speech-to-SL translator

Table 3. Set of words used for the Speech-to-SL translator

Abrazo	Abuelo	Adiós	Agua	Ayuda
Alegre	Ayer	Azul	Baño	Blanco
Caliente	Cama	Cariño	Casa	Comer
Cumpleaños	Descansar	Despedir	Despertar	Dinero
Enojo	Enfermar	Escuela	Estudiante	Feliz
Frio	Gracias	Gato	Gallina	Gustar
Habitación	Hambre	Hermano	Hijo	Hola
Hoy	Hospital	Importante	Invierno	Invitar
Leche	Leer	Libro	Luna	Mamá
Mañana	Médico	Mercado	Mesa	Mucho
Nariz	Niño	Noche	Papá	Poco
Querer	Rápido	Refresco	Reír	Romper
Salud	Señorita	Silla	Sopa	Sorpresa
Temor	Todo	Trabajar	Triste	Viajar

If a recognized word belongs to the vocabulary shown in Table 3, then the associated MSL 3D models are shown in the right side of the GUI. If the word is not found in the word-based MSL database, then it is described as the sequence of MSL models of each of its forming characters (found in the character-based MSL database).

5 Performance Results

The translation interface was evaluated with the 10 speakers that were used to test the ASR system (see Table 2). The test material consisted of:

Table 4. Alphabet used for the Speech-to-SL translator

A	B	C	CH	D	E	F	G	H	I	J
K	L	LL	M	N	Ñ	O	P	Q	R	RR
S	T	U	V	W	X	Y	Z			

- a set of words in Table 3;
- a set of 20 randomly selected sentences from the DIELSEME [12] system. These sentences presented in Table 5 consisted of 129 words with a vocabulary (number of unique words) of 84 words, where 60 were not present in the word-based MSL database. With these sentences, translation of words not present in the word-based MSL database was performed by alphabet spelling.

The performance of the system on the test sessions is shown in Table 6. The measure of performance was *WAcc* as the correct recognition of a word is linked to the corresponding MSL by the translator. In total, the test set consisted of 70 (words in Table 3) + 129 (words in sentences of Table 5) = 199 words. As presented, the overall accuracy was of 96.2% on 1990 uttered words (199 × 10 speakers). Although this performance is slightly less than the one reported in Table 2, the number of words in the test set is significantly bigger (1990 > 820). This performance is within the ranges of human transcription (96% - 98%) [32]; hence it was considered that the achieved recognition rates were satisfactory.

6 Conclusions and Future Work

In this paper a 3D modeling of the Mexican Sign Language (MSL) and its use within a Speech-to-LS translation system were presented. The use of Kinect as three-dimensional capture sensor and a particle filter for tracking finger configurations were explored as means to create accurate 3D models for the MSL. The models presented with a 3D avatar showed more details of the MSL performances than videos taken from a well known MSL dictionary. This modeling procedure can be applied to model other sign languages or be adapted for real-time sign recognition.

Table 5. Set of sentences used for the Speech-to-SL translator

1	mi **hermano** me dio un **abrazo** con **cariño**
2	el **gato** camina por el techo de la casa
3	olí la comida de la cocina y me dio **hambre**
4	**hoy** es el **cumpleaños** de mi **hijo**
5	mi **papá** y mi **mamá** van a salir de viaje
6	la máquina del coche está bien
7	el verano pasado llovió **mucho**
8	en el **hospital** hay **muchos médicos**
9	mi **hijo** me dio un regalo **sorpresa**
10	el siempre vive muy solitario
11	el payaso me hizo **reír**
12	el león caza para **comer**
13	ya es la hora de la **comida**
14	voy a **leer** el **libro** en mi cuarto
15	los **abuelos** quieren a sus nietos
16	la **niña** se sirvió **refresco**
17	los pájaros son de colores bonitos
18	el niño bañó al perro
19	esa **señorita** es soltera
20	la **sopa** está muy sabrosa

For this work, 100 3D models were created (70 words, 30 letters), and for a Speech-to-SL translator, two methodologies were implemented: (1) word-based and (2) character-based translation (for cases where no word-based models were available). For this system, accurate translation relies on the performance of an Automatic Speech Recognition (ASR) system. Thus, we also presented the details of design of an ASR for the Mexican Spanish language. The ASR system was extended compared to a previous prototype in order to improve such elements as the training speech corpus, the control of the language model, static and dynamic speaker adaptation, and the size of the test vocabulary. The ASR system achieved overall recognition rates of 96.2% for different test speakers. This performance is within ranges of human recognition.

As future work, we plan

- to build 3D models for all the vocabulary in the MSL (the DIELSEME dictionary has 608 signs associated to 535 spoken words) [12],

Table 6. Performance of the Speech-to-MSL translator when tested by 10 randomly selected users

Age	Gender	Origin	Occupation	N	D	S	I	% WAcc
23	Male	Oaxaca	Student	199	1	8	1	95.00
36	Female	Mexico City	Teacher	199	1	1	2	98.00
28	Male	Oaxaca	Student	199	1	6	0	96.50
32	Male	Puebla	Teacher	199	2	4	1	96.50
27	Male	Oaxaca	Student	199	1	9	2	94.00
21	Male	Mexico City	Student	199	0	7	1	96.00
21	Female	Puebla	Student	199	2	4	0	97.00
20	Male	Oaxaca	Student	199	1	8	0	95.50
35	Female	Mexico City	Teacher	199	0	6	1	96.50
19	Male	Oaxaca	Student	199	0	5	0	97.50
				1990	9	58	8	96.20

this also considers modeling of facial expressions;

- to improve the proposed system by integrating complex grammatical and syntactical rules for advanced Speech-to-MSL translation;
- to extend the system to capture signs in real time, thus allowing MSL-to-Speech/Text translation. This would support communication between deaf and hearing people, so tests with the participation of deaf people must be performed. In addition, advanced grammatical and syntactical rules are to be integrated to accomplish accurate MSL-to-Speech translation.

References

1. **Nuance Communications, Inc. (2012).** Dragon Speech Recognition Software. Retrieved from http://www.nuance.com/dragon/index.htm.
2. **IBM (2012).** WebSphere Voice. Retrieved from http://www-01.ibm.com/software/voice/.
3. **Lavie, A., Waibel, A., Levin, L., Finke, M., Gates, D., Gavalda, M., Zeppenfeld, T., & Zhan, P. (1997).** JANUS III: Speech-To-Speech Translation In Multiple Languages. *IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-97)*, Munich, Germany, 1, 99–102.
4. **Parker, M., Cunningham, S., Enderby, P., Hawley, M., & Green, P. (2006).** Automatic speech recognition and training for severely dysarthric users of assistive technology: The STARDUST project. *Clinical Linguistics and Phonetics*, 20(2-3), 149–156.
5. **Dalby, J. & Kewley-Port, D. (1999).** Explicit Pronunciation Training Using Automatic Speech Recognition Technology. *Computer-Assisted Language Instruction Consortium (CALICO) Journal*, 16(3), 425–445.
6. **Rosetta Stone (2012).** Rosetta Stone Version 4 TOTALe. Retrieved from http://www.rosettastone.com/learn-spanish.
7. **Cox, S., Lincoln, M., Nakisa, M., Wells, M., Tutt, M., & Abbott, S. (2003).** The Development and Evaluation of a Speech to Sign Translation System to Assist Transactions. *International Journal of Human Computer Interaction*, 16(2), 141–161.
8. **San-Segundo, R., Barra, R., D'Haro, L.F., Montero, J.M., Córdoba, R., & Ferreiros, J. (2006).** A spanish speech to sign language translation system for assisting deaf-mute people. *Ninth International Conference on Spoken Language Processing (INTERSPEECH 2006-ICSLP),* Pittsburgh, PA, USA, 1399–1402.
9. **Baldassarri, S., Cerezo, E., & Royo-Santas, F. (2009).** Automatic Translation System to Spanish Sign Language with a Virtual Interpreter. *Human-Computer Interaction - INTERACT 2009*, *Lecture Notes in Computer Science*, 5726, 196–199.
10. **López-Colino, F. & Colás, J. (2011).** The Synthesis of LSE Classifiers: From Representation

to Evaluation. *Journal of Universal Computer Science*, 17(3), 399–425.

11. **Massó, G. & Badia, T. (2010).** Dealing with Sign Language Morphemes in Statistical Machine Translation. *4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*, Valletta, Malta, 154–157.
12. **Calvo, M.T. (2004).** Diccionario Español - Lengua de Señas Mexicana (DIELSEME): estudio introductorio. Dirección de Educación Especial: México.
13. **Saldivar-Piñon, L., Chacon-Murguia, M., Sandoval-Rodriguez, R., & Vega-Pineda, J. (2012).** Human Sign Recognition for Robot Manipulation. *Pattern Recognition*, *Lecture Notes in Computer Science*, 7329, 107–116.
14. **Rios, D. & Schaeffer, S. (2012).** A Tool for Hand-Sign Recognition. *Pattern Recognition*, *Lecture Notes in Computer Science*, 7329, 137–146.
15. **Clymer, E., Geigel, J., Behm, G., & Masters, K. (2012).** *Use of Signing Avatars to Enhance Direct Communication Support for Deaf and Hard-of-Hearing Users*. National Technical Institute for the Deaf (NTID), Rochester Institute of Technology, United States.
16. **Microsoft Co. (2012).** Kinect for Windows. Retrieved from http://www.microsoft.com/en-us/kinectforwindows/.
17. **Albrecht, I., Haber, J., & Seidel, H.P. (2003).** Construction and Animation of Anatomically Based Human Hand Models. *2003 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, San Diego, CA, USA, 98–109.
18. **Bretzner, L., Laptev, I., & Lindeberg, T. (2002).** Hand gesture recognition using multi-scale colour features, hierarchical models and particle filtering. *Fifth IEEE International Conference on Automatic Face and Gesture Recognition*, Washington, DC, USA, 423–428.
19. **Oikonomidis, I., Kyriazis, N., & Argyros, A.A. (2011).** Efficient model-based 3D tracking of hand articulations using Kinect. *Proceedings of the British Machine Vision Conference (BMVC 2011)*, Dundee, UK, (101.1–101.11).
20. **Trigo, T.R. & Pellegrino, S.R. (2010).** An analysis of features for hand-gesture classification. *17th International Conference on Systems, Signals and Image Processing (IWSSIP 2010)*, Rio de Janeiro, Brazil, 412–415.
21. **DAZ Productions (2012).** DAZ Studio 4.5. Retrieved from http://www.daz3d.com/daz-studio-4-pro/.
22. **Bonilla, G. (2012).** *Interfaz de Voz para Personas con Disartria*. Tesis Ingeniero en Computación, Universidad Tecnológica de la Mixteca (UTM), Huajuapan, Oaxaca, Mexico.
23. **Jurafsky, D. & Martin, J.H. (2009).** *Speech and Language Processing: an introduction to natural language processing, computational linguistics, and speech recognition*. Upper Saddle River, N.J.: Pearson Prentice Hall.
24. **Young, S., Evermann, G., Gales, M., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev., V & Woodland, P. (2006).** *The HTK Book (for HTK Version 3.4)*. Cambridge University Engineering Department: Cambridge, UK.
25. **Leggetter, C.J. & Woodland, P.C. (1995).** Maximum likelihood linear regression for speaker adaptation of continuous density hidden Markov models. *Computer Speech and Language*, 9(2), 171–185.
26. **Cuétara, J.O. (2004).** *Fonética de la Ciudad de México. Aportaciones desde las tecnologías del habla.* Maestro en Lingüistica Aplicada, Universidad Nacional Autónoma de México (UNAM), México, D.F.
27. **Pineda, L.A., Villaseñor, L., Cuétara, J., Castellanos, H., & López, I. (2004).** DIMEx100: A new phonetic and speech corpus for Mexican Spanish. *Advances in Artificial Intelligence (IBERAMIA 2004)*, *Lecture Notes in Computer Science*, 3315, 974–983.
28. **Pineda, L.A., Castellanos, H., Cuétara, J., Galescu, L., Juárez, J., Llisterri, J., Pérez, P., & Villaseñor, L. (2010).** The corpus dimex100: Transcription and evaluation. *Language Resources and Evaluation*, 44(4), 347–370.
29. **Trujillo-Romero, F. & Caballero-Morales, S.O. (2012).** Towards the Development of a Mexican Speech-to-Sign-Language Translator for the Deaf Community. *Acta Universitaria*, 22(NE-1), 83–89.
30. **Sjolander, K. & Beskow, J. (2006).** Wavesurfer. Retrieved from http://www.speech.kth.se/wavesurfer/.
31. **Rabiner, L. (1989).** A tutorial on hidden Markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2), 257–286.
32. **National Institute of Standards and Technology (NIST) (s.f.).** The History of Automatic Speech Recognition Evaluations at NIST. Retrieved from

http://www.itl.nist.gov/iad/mig/publications/ASRhistory/.

Santiago-Omar Caballero-Morales received his Ph.D. in Computer Science by the University of East Anglia (UEA) in the United Kingdom in 2009. Currently he is a Full-Time Professor-Researcher in the Postgraduate Division at the Technological University of the Mixteca (UTM) located in Huajuapan de Leon, Oaxaca (Mexico). His research interests are automatic speech recognition for normal and disordered speech, human-robot interaction, speech translation for native languages, combinatorial optimization, statistical quality control, and industrial simulation.

Felipe Trujillo-Romero received his Ph.D. in Informatics Systems by the Institut National Polytechnique de Toulouse France in 2008. He is currently a Professor-Researcher at the Technological University of the Mixteca (UTM) located in Huajuapan de Leon, Oaxaca (Mexico). His current research interests include evolutionary algorithms, robotics, parallel algorithms, and computer vision.

Article received on 15/10/2012; accepted 21/06/2013.