Towards the Development of a Mexican Speech-to-Sign-Language Translator for the Deaf Community

Felipe Trujillo-Romero* , Santiago-Omar Caballero-Morales*

ABSTRACT

A significant population of Mexican people are deaf. This disorder restricts their social interaction skills with people who don't have such disorder and viceversa. In this paper we present our advances towards the development of a Mexican Speech-to-Sign-Language translator to assist normal people to interact with deaf people. The proposed design methodology considers limited resources for (1) the development of the Mexican Automatic Speech Recogniser (ASR) system, which is the main module in the translator, and (2) the Mexican Sign Language (MSL) vocabulary available to represent the decoded speech. Speech-to-MSL translation was accomplished with an accuracy level over 97% for test speakers different from those selected for ASR training.

RESUMEN

Una parte significativa de la población mexicana es sorda. Esta discapacidad restringe sus habilidades de interacción social con personas que no tienen dicha discapacidad y viceversa. En este artículo presentamos nuestros avances hacia el desarrollo de un traductor Voz-a-Lenguaje-de-Señas del español mexicano para asistir a personas sin discapacidad a interactuar con personas sordas. La metodología de diseño propuesta considera limitados recursos para (1) el desarrollo del Reconocedor Automático del Habla (RAH) mexicano, el cual es el módulo principal del traductor, y (2) el vocabulario del Lenguaje de Señas Mexicano (LSM) disponible para representar las oraciones reconocidas. La traducción Voz-a-Lenguaje-de-Señas fue lograda con un nivel de precisión mayor al 97% para usuarios de prueba diferentes de aquellos seleccionados para el entrenamiento del RAH.

Recibido: 6 de Enero del 2012
Aceptado: 14 de Febrero del 2012

INTRODUCTION

Development of assistive technology for deaf people has been made for different contexts of use. In [1] Speech-to-Spanish Sign Language (Lengua de Signos Española, LSE) translation was developed for sentences spoken by an official when assisting people applying for, or renewing their Identity Card in Spain. Another system, called SiSi (Say It Sign It) [2] was developed for more flexible Speech-to-Sign Language translation (in this case, translation to the British Sign Language, BSL).

Such systems required intensive research in language modelling, in both, spoken and sign forms. In the case of Spanish, besides the study in [1], there has been research in [3] related to the statistical translation of an ASR's output (i.e., Speech-to-Text translation) into LSE.

Another approach was presented in [4] where the Spanish Speech-to-Sign translation system considered the morphological and syntactical relationships of words in addition to the semantics of their meaning in the Spanish language.The work of Massó and Badia [5] used a morpho-syntactic approach to generate a statistical translation machine for the Catalán language.All these Speech-to-Sign translation systems made use of a 3D avatar to perform the sign representations of recognised spoken words. Although there is research in the development of such translation systems for the Spanish language, there is not significant work towards the development of a translator for the Mexican Spanish language.

Palabras clave:
Tecnología asistiva; traducción de voz a señas; lenguaje mexicano de señas; reconocimiento automático del habla.
Keywords:
Assistive technology; speech-to-sign translation ; mexican sign language; automatic speech recognition.

*División de Posgrado, Universidad Tecnologica de la Mixteca, Carretera a Acatlima Km 2.5, C.P. 69000, Huajuapan de Leon, Oaxaca, Mexico. E-mail: ftrujillo@mixteco.utm.mx, scaballero@mixteco.utm.mx

Hence, in this paper we present our advances towards the development of a Mexican Speech-to-Mexican-Sign-Language (MSL) translation system. The proposed structure of this system is shown in figure 1, and the details about the design of each element are described in the following sections. The ASR engine, trained with few but representative speakers achieved recognition accuracy of MSL vocabulary words of 97.2%. Hence, the structure of this paper is as follows: in Section *Automatic Speech Recognition Module* the details of the multi-user ASR system for the Mexican spanish language are shown; in Section *Text Interpreter and MSL Database* the details about the structure of the Speech-to-Sign Language translator are presented (i.e., the text interpreter); in Section *Performance Results* the performance results of the integrated interface are shown; finally, in Section *Conclusions and Future Work* the conclusions and future plans for this project are discussed.

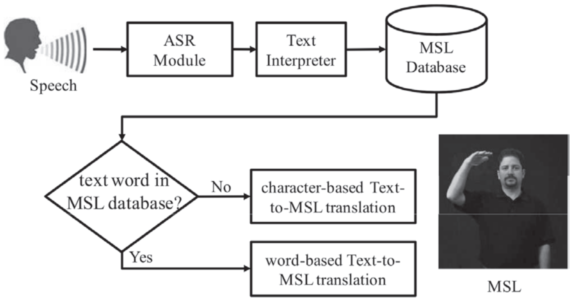

Speech
ASR Module
Text Interpreter
MSL Database
text word in MSL database?
No
Yes
character-based Text-to-MSL translation
word-based Text-to-MSL translation
MSL

Figure 1 . Structure of the Speech-to-MSL translator.

AUTOMATIC SPEECH RECOGNITION MODULE

In order to perform reliable speech-to-sign translation, speech must be decoded (recognised) accurately. A robust ASR system can perform such task. There are different techniques such as Artificial Neural Networks (ANNs [6]), Hidden Markov Models (HMMs [7]), Weighted Finite State Transducers (WFSTs [8]), etc., to build the functional components of the ASR module for the translation system.

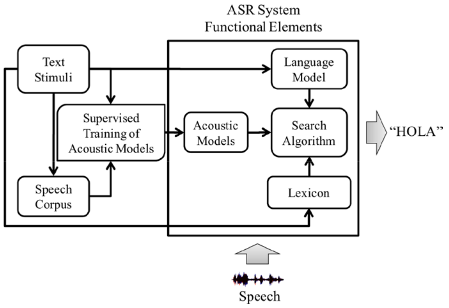


Figure 2 . Structure of an *ASR system* .

In figure 2 the standard estructure of an ASR system is shown, and each component is explained in the following sections.

Training Speech Corpus

To accomplish robust ASR performance, the system must be trained with a wide variety of speech patterns, and currently there are large databases of speech data, known as Speech Corpora (i.e., WSJ [9], TIMIT [10], etc.), available for this purpose. For the mexican spanish language (or latin american spanish) there are few of these resources. The most significant is the Mexican Spanish Corpus DIMEx100 developed at the Instituto de Investigaciones en Matemáticas Aplicadas y Sistemas ("Applied Mathematics and Systems Research Institute") IIMAS of the National Autonomous University of Mexico UNAM [11, 12].

Due to licensing procedures still in process to make available this resource for distribution for other projects, we were unable to use this corpus for the supervised training of the ASR module. Thus, we decided to explore the situation of training this module with limited speech data, and measure the highest level of accuracy achievable with the resulting module when tested by different speakers.

It was assumed that the robustness of the ASR system trained with few speech data could be accomplished if:

1. The training speakers were representative of the main speech features in a language.
2. There were enough speech samples for acoustic modelling.
3. The vocabulary of the application were not large (< 1000 words).
4. Dynamic speaker adaptation were performed while using the system.

To get the speech samples, six speakers were recruited based in the following criteria:

1. Place of origin close to the central region of Mexico (in this case, Mexico City and Puebla).
2. Age within the 15 - 60 years range.
3. Genre (equal number of male and female participants).

In table 1 the details of the six participants are shown.

Table 1 .

Training speakers for the ASR system.

Male Speakers	**S1**	**S2**	**S3**
Age	17	55	27
Origin	Mexico City	Oaxaca	Puebla
Female Speakers	**S4**	**S5**	**S6**
Age	37	15	50
Origin	Mexico City	Oaxaca	Puebla

A text stimuli was selected for purposes of speech recording as the speech samples must be phonetically balanced (i.e., all phonemes in the mexican language must be present in the corpus). This text is read by the participants and their speech is then recorded. The stimuli text consisted of: (1) 49 words with the form consonant-vowel-consonant; (2) a short story taken from a narrative; and (3) 16 short sentences designed to further include all phonemes in the mexican spanish language. The definition of phonemes for the mexican spanish language was obtained with the tool TranscribeMex [12] which was developed to phonetically label the DIMEX corpus. TrancribeMex was designed to define the sequences of phonemes that form a word considering the standard pronunciation of people in Mexico City [13, 11]. This was the reason to recruit speakers from (or very close to) this region. In table 2 the mexican phonemes and their number of occurrences (frequency) in the stimuli text are shown.

Table 2 .

Frequency of mexican phonemes in the stimuli text.

	Phoneme	**Frequency**		**Phoneme**	**Frequency**
1	/a/	183	15	/o/	95
2	/b/	44	16	/p/	231
3	/tS/	18	17	/r/	10
4	/d/	34	18	/r(/	94
5	/e/	121	19	/s/	69
6	/f/	17	20	/t/	45
7	/g/	19	21	/u/	41
8	/i/	76	22	/ks/	10
9	/x/	11	23	/Z/	12
10	/k/	46	24	/_D/	10
11	/l/	44	25	/_G/	6
12	/m/	33	26	/_N/	76
13	/n/	28	27	/_R/	44
14	/ñ/	16	28	/sil/	410

The speech corpus was recorded[1] in the following way for each speaker: the set of 49 words was read five times, the short story was read three times, and the 16 sentences were read once. These speech samples must be labelled at the orthographic and phonetic levels to perform supervised training of the acoustic models of the ASR system. Orthographic labelling was performed manually with the software Wavesurfer [14], and with the phonemes definitions obtained with TrancribeMex these labels were decomposed into phoneme labels. In figure 3 an example of these labels is shown.

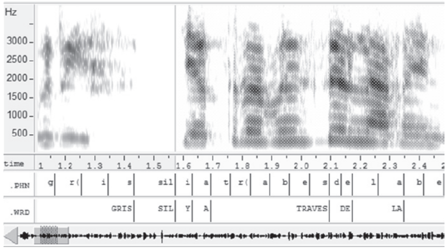


Figure 3 . Orthographic (.WRD) and phonetic (.PHN) labelling of speech data with Wavesurfer.

After the training speech corpus was finished we proceeded to build the functional elements of the ASR system shown in figure 2. The HTK library [15] was used for this purpose.

Functional Elements
Acoustic Models

The technique used for acoustic modelling was HMMs [7], and the implementation tool was HTK [15]. In figure 4 the structure of the HMMs used for acoustic modelling of phonemes is shown. This is a standard three-state left-to-right architecture with eight mixture gaussian components per state [16, 15]. For supervised training, the speech corpus was coded into Mel Frequency Cepstral Coefficients (MFCC's). The front-end used 12 MFCC's plus energy, delta, and acceleration coefficients [15].

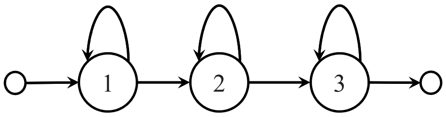


Figure 4 . Three-state left-to-righ structure of a HMM for phoneme acoustic modelling.

The supervised training of each phoneme's HMM (28 in total) was performed with the MFCC coded speech corpus, together with its phonetic labels, by means of the *HInit* (for HMM initialization) and *HRest / HERest* (for HMM re-estimation) HTK utilities[2].

[1] The speech was recorded with a Sony lcd-bx800 recorder with a sampling frequency of 8 kHz monoaural in WAV format.

[2] These utilities estimate the parameters of the HMMs by performing temporal re-alignment of the speech data with their respective phonetic labels using the Baum-Welch and Viterbi algorithms [16, 15].

Language Model and Lexicon

The Language Model (LM) represents a set of rules or probabilities that restricts the recognised sequence of words from the ASR system to valid sequences. Thus, this element guides the search (decoding) algorithm to find the most likely sequence of words that best represent an input speech signal. Commonly, N-grams are used for the LM, and for this work, bigrams (N=2)were used for continuous speech recognition [16, 15]. Estimation of bigrams was performed with the *HLStats* and *HBuild* HTK utilities. *HLStats* estimates the frequency of each single word and pairs of words in the text stimuli, and *HBuild* constructs with that information a network for word recognition.

The Lexicon specifies the sequences of phonemes that form each word in the application's vocabulary. This element was developed while the speech corpus was being phonetically labelled (see Section *Automatic Speech Recognition Module-Training Speech Corpus*).

Search Algorithm

The Viterbi algorithm is widely used for speech recognition [16]. This task consists in finding (searching) the sequence of words that best match the speech signal. Viterbi decoding was implemented with the utility *HVite* of HTK.

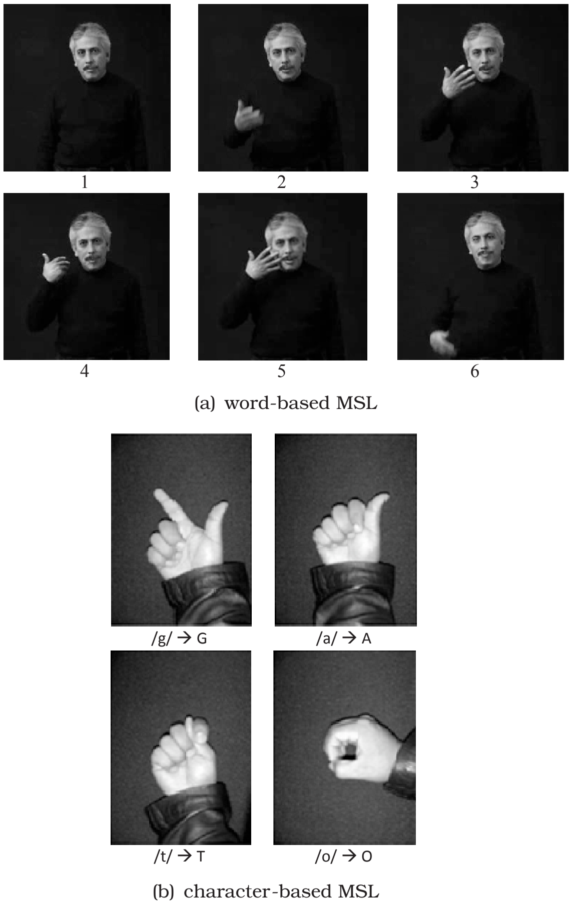
Speaker Adaptation

Commercial ASR systems are trained with hundreds or thousands of speech samples from different speakers. When a new user wants to use such system, it is common to ask the user to read some words or narratives to provide speech samples that will be used by the system to adapt its acoustic models to the patterns of the user's voice. Commercial ASR systems are robust enough to get benefits by the implementation of adaptation techniques such as MAP or MLLR [15, 17]. For this work, a large corpus was not available, and thus, the ASR system was trained with speech samples from six speakers (see table 1).

Maximum Likelihood Linear Regression (MLLR) [17] was the adaptation technique used for the ASR system in order to make it usable for other speakers. For this task, the 16 balanced sentences (see Section *Automatic Speech Recognition Module-Training Speech Corpus*) were used as stimuli. This technique is based on the assumption that a set of linear transformations can be used to reduce the mismatch between an initial HMM model set and the adaptation data. In this work, these transformations were applied to the mean and variance parameters of the gaussian mixtures of the HMM's of the ASR system. A regression class tree with 32 terminal nodes was used for the dynamic implementation of the MLLR adaptation [15, 17].

TEXT INTERPRETER AND MSL DATABASE

The text interpreter searches in a *MSL Database* (see figure 1) the MSL representation that best matches the recognised (decoded) speech. If the recognised word is found in the MSL database, then the interpreter proceeds to display the sequence of MSL movements associated to that word. Otherwise, if the word is not found in the database, the word is "spelled", and the word is described with the MSL representations associated to each letter (character) that form the word. This was accomplished by decomposing the word into phonemes with TranscribeMex and then assigning to each phoneme an alphabet character in the MSL vocabulary (see Section *Text Interpreter and MSL Database-MSL Vocabulary*).



(a) word-based MSL

(b) character-based MSL

Figure 5 . MSL representations of the word GATO.

Hence, the MSL database consists of animated representations of MSL movements that describe the mexican spanish vocabulary. The word-based MSL representations were taken from the video library of the *DIELSEME* [18] system. These videos, in *SWF* format, were converted into *AVI* format [3] with the software *AVS Video Converter ver. 7.1.2.480*. The character-based MSL representations were performed by a MSL signer and stored as pictures in *JPG* format.

MSL Vocabulary

The vocabulary used by the interface is shown in table 3. The main vocabulary consists of 25 words for the word-based Text-to-MSL translation. If a recognised word is not within this set, then it is described in terms of the alphabet characters that form the word. For this task, a set of 23 characters was considered for the character-based Text-to-MSL translation. Note that the movements that are performed to describe a word in MSL are not equivalent to the sequence of character-based MSL movements. Figure 5(a) presents the MSL representation of the word *GATO* (cat), and figure 5(b) the character-based representation of the same word. Note that both representations differ from each other.

The character-based MSL is proposed as an alternative to flexible communication for large vocabularies without the need to animate each word in the mexican language.

Table 3 .

MSL vocabulary.

Words		**Alphabet**	
Hola	Hijo	A	P
Adios	Niño	B	Q
Hoy	Hermano	C	R
Ayer	Blanco	D	S
Mañana	Rojo	E	T
Noche	Azul	F	U
Alegre	Casa	G	V
Feliz	Silla	H	W
Triste	Mesa	I	X
Temor	Cama	L	Y
Enojo	Habitación	M	
Mamá	Gracias	N	
Papá		O	

PERFORMANCE RESULTS

The multi-user ASR module together with the Text Interpreter/MSL Database and the video animations were integrated within a graphical interface for its use by test speakers. Thus, the Speech-to-MSL interface is shown in figure 6.

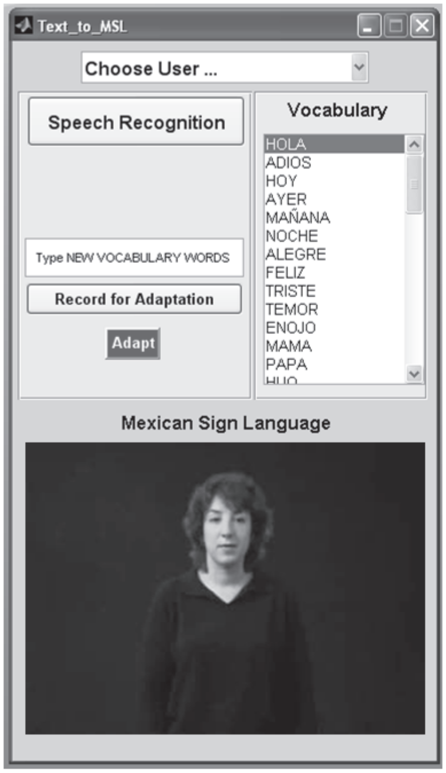


Figure 6 . Speech-to-MSL Translator.

In the field *Choose User ...* the user can type his/her name or select an existing user already registered in the system. By doing this, the interface automatically creates the files needed to adapt the system to the new user, or to load the user's adapted acoustic models to perform speech recognition. If the user is already registered then he/she can proceed to use the Text-to-MSL translator by pressing the button *Speech Recognition*, otherwise the user must proceed to adapt the system. This is accomplished by entering text stimuli (i.e., adaptation sentences, see Section *Automatic Speech Recognition Module, Training Speech Corpus*) in the field *Type NEW VOCABULARY WORDS*, and pressing *Record for Adaptation* to record the user's speech for that stimuli. The user can enter any text and record as many words as desired. After the adaptation data is recorded the user just needs to press *Adapt* to execute the interface's MLLR adaptation process. Note that this task is cumulative, thus the adaptation speech data is stored within the interface. An existing user can add more vocabulary and further improve the performance of his/her adapted

[3] Intel Indeo Video 3.2 codec.

acoustic models. This was considered as "dynamic" speaker adaptation. All the additional text/vocabulary is updated in the ASR's language model and lexicon(See section Automatic Speech Recognition Module, Functional Elements).

Tests were performed with ten users. Prior to use the Speech-to-MSL translator the test users were registered, and adaptation was performed with a stimuli text of 16 phonetically balanced sentences (see Section Automatic Speech Recognition Module,Training Speech Corpus). The metric used to measure the performance of the Speech-to-MSL translator was the Word Error Rate (WER) which is computed as:

$$WER = 1 - \frac{N - D - S - I}{N} \qquad (1)$$

where D, S, and I are deletion, substitution, and insertion errors in the recognised speech (text output of the ASR module) which affect the MSL translation. N is the number of words in the correct ASR's output. The translation system was tested with ten speakers and the 25 words in the main MSL vocabulary as stimuli. Besides these words, 15 were added to the system to test character-based MSL translation and dynamic vocabulary construction. The stimuli was read (spoken) just once, and the first result generated by the translator was considered as the definitive output. The performance results are presented in table 4. In total a WER of 2.8% was achived by the system, which is equivalent to a recognition word accuracy of 97.2%. Considering that the WER for human transcription is within the range of 2%-4%, and ASR performance for read text is within the range of 3.5%-20% for vocabularies < 1,000 words [19], the performance of this system for the MSL vocabulary is comparable to that of human perception and other systems for small vocabulary. The word-based and character-based MSL animations for words in the MSL database were performed smoothly.

Table 4 .

Performance of the Speech-to-MSL Translator.

Test Speakers	N	D,I,S Errors	WER
S1	40	1	2.5%
S2	40	1	2.5%
S3	40	0	0.0%
S4	40	2	5.0%
S5	40	0	0.0%
S6	40	3	7.5%
S7	40	0	0.0%
S8	40	3	7.5%
S9	40	0	0.0%
S10	40	1	2.5%
Total	400	11	2.8%

CONCLUSIONS AND FUTURE WORK

In this paper the advances towards the development of a Mexican Speech-to-MSL translator were presented. Even with limited resources, multi-user ASR performance of 97.2% was achieved in test sessions of 400 words in total.

Although at this stage the MSL vocabulary is small, the results reported here give confidence about the feasibility of the project and the levels of performance that the system can achieve. However we realise that much work is needed and as future work the following points are considered:

- Improve the Speech-to-MSL translator and the interface to control the influence of the language model over the recognition procedure;
- Obtain a more extensive view of the performance of the ASR system when testing the system with a larger vocabulary;
- Increase the animated database of the MSL vocabulary: Kinect is being considered to be used as a tool for motion capture to map physical MSL representations to an animated 3D avatar for the translation system;
- Allow translation of continuous speech (sentences) into MSL considering grammar and syntactical rules;
- Develop the complementary translation system: MSL-to-Speech translation.

REFERENCES

[1] 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. *Journal of Speech Communication* 50:1009--1020.

[2] Sys Consulting, IBM, RNID, University of East Anglia (2011). Say It Sign It: Converting speech to deaf signs. *http://www.sys-consulting.co.uk/web/Case_SISI.pdf.*

[3] López Alvarez, V., San Segundo, R., Martín Clemente, R., Lucas, J.M., Barra, R., Chicote, R. (2010). Estudio del tipo de alineamiento en un sistema de traducción estadística del castellano a lengua de signos española (lse). *Procesamiento del Lenguaje Natural* 45:207--214.

[4] Baldassarri, S., Cerezo, E., Royo-Santas, F. (2009). Automatic Translation System to Spanish Sign Language with a Virtual Interpreter. In *Proc. of Human-Computer Interaction - INTERACT 2009, Lecture Notes in Computer Science, Springer Berlin,* (5726)196-199.

[5] Massó, G., Badia, T. (2010). Dealing with Sign Language Morphemes in Statistical Machine Translation. In Proc. of the *4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies* , 154-157.

[6] Jayaram, G., Abdelhamied, K. (1995). Experiments in dysarthric speech recognition using artificial neural networks. *Journal of Rehabilitation Research and Development* 42:162-169.

[7] Rabiner, L. (1989). A tutorial on Hidden Markov Models and selected applications in speech recognition. In Proc. IEEE, 37:257-286.

[8] Mohri, M., Pereira, F., Riley, L. (2002). Weighted finite state transducers in speech recognition. *Computer Speech and Language* 16:69-88.

[9] Robinson,T. (1995). WSJCAM0: A British English Speech Corpus for Large Vocabulary Continuous Speech Recognition. In Proc. *IEEE Conf. on Acoustics, Speech and Signal Processing* , 81-84.

[10] Byrd, D. (1992). Preliminary results on speaker-dependent variation in the *TIMIT* database. J. Acoust. Soc. Amer.

[11] Pineda, L.A., Villaseñor, L., Cuétara, J., Castellanos, H., López, I. (2004). DIMEx100: A new phonetic and speech corpus for Mexican Spanish. In *Advances in Artificial Intelligence* , Iberamia-2004.

[12] Pineda, L.A., Villaseñor, L., Cuétara, J., Castellanos, H., Galescu, L., Juárez, J., Llisterri, J., Pérez, P. (2010). The corpus dimex100: Transcription and evaluation. *Language Resources and Evaluation* 44:347-370.

[13] Cuétara, J. (2004). Fonética de la Ciudad de México: *Aportaciones desde las tecnologías del habla* . MSc. Dissertation, National Autonomous University of Mexico (UNAM), Mexico.

[14] Beskow, J., Sjolander, K. (2011). Wavesurfer v.1.8.8.3p3. http://www.speech.kth.se/wavesurfer/.

[15] Young, S., Woodland, P. (2006). The HTK Book (for HTK Version 3.4). Cambridge University Engineering Department.

[16] Jurafsky, D., Martin, J.H. (2009). *Speech and Language Processing* . Pearson: Prentice Hall.

[17] 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.

[18] Secretaria de Educación Pública (2004). *Diccionario Español - Lengua de Señas Mexicana (DIELSEME): estudio introductorio.* Dirección de Estudios Especiales, Mexico.

[19] National Institute of Standards and Technology (NIST) (2011). The History of Automatic Speech Recognition Evaluations at NIST. *http://www.itl.nist.gov/iad/mig/publications/ASRhistory/ index.html.*