Expert Systems with Applications 40 (2013) 1283–1295

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

LSESpeak: A spoken language generator for Deaf people

Verónica López-Ludeña *, Roberto Barra-Chicote, Syaheerah Lutfi, Juan Manuel Montero, Rubén San-Segundo

Departamento de Ingeniería Electrónica, ETSI Telecomunicación, Universidad Politécnica de Madrid, Spain

ARTICLE INFO

Keywords:
Spanish sign language
LSE: Lengua de Signos Española
SMS to Spanish translation
Speech generation from LSE
Emotional speech synthesis

ABSTRACT

This paper describes the development of LSESpeak, a spoken Spanish generator for Deaf people. This system integrates two main tools: a sign language into speech translation system and an SMS (Short Message Service) into speech translation system. The first tool is made up of three modules: an advanced visual interface (where a deaf person can specify a sequence of signs), a language translator (for generating the sequence of words in Spanish), and finally, an emotional text to speech (TTS) converter to generate spoken Spanish. The visual interface allows a sign sequence to be defined using several utilities. The emotional TTS converter is based on Hidden Semi-Markov Models (HSMMs) permitting voice gender, type of emotion, and emotional strength to be controlled. The second tool is made up of an SMS message editor, a language translator and the same emotional text to speech converter. Both translation tools use a phrase-based translation strategy where translation and target language models are trained from parallel corpora. In the experiments carried out to evaluate the translation performance, the sign language-speech translation system reported a 96.45 BLEU and the SMS-speech system a 44.36 BLEU in a specific domain: the renewal of the Identity Document and Driving License. In the evaluation of the emotional TTS, it is important to highlight the improvement in the naturalness thanks to the morpho-syntactic features, and the high flexibility provided by HSMMs when generating different emotional strengths.

1. Introduction

In the world, there are around 70 million people with hearing deficiencies (information from World Federation of the Deaf). Deafness brings about significant communication problems: deaf people cannot hear and most of them are unable to use written languages, having serious problems when expressing themselves in these languages or understanding written texts. They have problems with verb tenses, concordances of gender and number, etc., and they have difficulties when creating a mental image of abstract concepts. This fact can cause deaf people to have problems when accessing information, education, job, social relationship, culture, etc. Deaf people use a sign language (their mother tongue) for communicating and there are not enough sign-language interpreters and communication systems. In the USA, there are 650,000 Deaf people (who use a sign language), although there are more people with hearing deficiencies, but only 7000 sign-language interpreters, i.e., a ratio of 93 deaf people to 1 interpreter. In Finland we can find the best ratio, 6–1, and in Slovakia the worst with 3000 users to 1 interpreter (Wheatley and Pabsch, 2010). In Spain this ratio is 221–1. This information shows the need to develop automatic translation systems with new technologies for helping hearing and deaf people to communicate between themselves.

It is necessary to make a difference between "deaf" and "Deaf": the first one refers to non-hearing people, and the second one refers to people who use a sign language as the first way to communicate being part of the "Deaf community". Each country has a different sign language, but there may even be different sign languages in different regions in the same country. There is also an international sign language, but most of deaf people do not know it. However, national sign languages are fully-fledged languages that have a grammar and lexicon just like any spoken language, contrary to what most people think. Traditionally, deafness has been associated to people with learning problems but this is not true. The use of sign languages defines the Deaf as a linguistic minority, with learning skills, cultural and group rights similar to other minority language communities.

According to information from INE (Statistic Spanish Institute), there are 1,064,000 deaf people in Spain and 50% are more than 65 years old. They are a geographically dispersed population, producing more social isolation. 47% of deaf population do not have basic studies or are illiterate, and only between 1% and 3% have finished their studies (as opposed to 21% of Spanish hearing people). Also, 20% of the deaf population is unemployed (30% for women).

* Corresponding author. Address: Grupo de Tecnología del Habla, Dpto. Ingeniería Electrónica, ETSI Telecomunicación, Universidad Politécnica de Madrid, Ciudad Universitaria s/n, 28040-Madrid, Spain.
E-mail address: veronicalopez@die.upm.es (V. López-Ludeña).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.08.062

According to the information presented above, deaf people are more vulnerable and they do not have the same opportunities as hearing people. They cannot access information and communication in the same way as hearing people do: TV programs, multimedia content on the internet and personal public services. All these aspects support the need to generate new technologies in order to develop automatic translation systems for converting this information into sign language. This paper presents LSESpeak, a software application that integrates speech and language processing technologies for helping Deaf to generate spoken Spanish. This application is very useful for interacting, in a face to face communication, to hearing people that do not know LSE. For instance, this system could be used in specific domains like the renewal of the Identity Document and Driving License. Deaf person would carry a tablet with the integrated translation system that would help to interact with hearing employees. LSESpeak complements a Spanish into LSE translation system (San-Segundo et al., 2011), allowing a two direction interaction.

The paper is organised as follows. Section 2 reviews the state of art on related assistive systems and technologies. Section 3 describes an overview of LSESpeak. Section 4 presents the visual interface of LSESpeak. Section 5 evaluates the sign language into text translation utility. Section 6 details the tool for translating SMS into text. Section 7 describes the emotional text to speech converter. Finally, the main conclusions are described in section 8.

2. State of the art

In order to eliminate the communication barriers between deaf and hearing people, it is necessary not only to translate speech into sign language (San-Segundo et al., 2011) but also to generate spoken language from sign language, allowing a fluent dialogue in both directions.

In previous projects, such as VANESSA (http://www.visicast.cmp.uea.ac.uk/eSIGN/Vanessa.htm) Tryggvason (2004), this problem was solved by asking the Deaf to write down the sentence in English (or Spanish in our case) and then a text to speech (TTS) converter can generate the speech. But this is not a good solution because a very high percentage of Deaf people do not write properly in Spanish. Sign language is their first language, and their ability to write or understand written language may be poor in many cases. Because of this, a great deal of effort has been made in recognising sign language and translating it into spoken language by using a language translator and a TTS converter. The main efforts have focused on recognising signs from video processing. The systems developed so far are very person or environment dependent (Vogler & Metaxas, 2001), or they focus on the recognition of isolated signs (von Agris, Schneider, Zieren, & Kraiss, 2006; Wang et al., 2006) which can often be characterised just by the direction of their movement. In (Yung-Hui and Cheng-Yue, 2009), authors propose a system for recognizing static gestures in Taiwan sign languages (TSL), using 3D data and neural networks trained to completion. In (Karami, Zanj and Sarkaleh, 2011) a system for recognizing static gestures of alphabets in Persian sign language (PSL) using Wavelet transform and neural networks is presented. 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 Munib, Habeeb, Takruri, and Al-Malik (2007). In Sylvie and Surendra (2005) a review of research into sign language and gesture recognition is presented.

In the Computer Science department of the RWTH Aachen University, P. Dreuw supervised by H. Ney is making a significant effort into recognizing continuous sign language from video processing (Dreuw & Deselaers, 2008; Dreuw, 2008; Dreuw, Stein, & Ney, 2009). The results obtained are very promising but they are not yet good enough to be accepted by the Deaf. Once the sign language is recognised, the sign sequence is translated into a word sequence which is then passed to a TTS converter to generate the speech. In Europe, the two main research projects that focus on sign language recognition are DICTA-SIGN (Efthimiou, Hanke, Bowden, Collet, & Goudenove, 2010; Hanke & Wagner, 2010) and SIGN-SPEAK (Dreuw et al., 2010a, 2010b), both financed by The European Commission within the Seventh Framework Program. DICTA-SIGN (http://www.dictasign.eu/) aims to develop the technologies necessary to make Web 2.0 interactions in sign language possible: users sign into to a webcam using a dictation style. The computer recognizes the signed phrases, converts them into an internal representation of sign language, and finally an animated avatar signs them back to the users. In SIGN-SPEAK (http://www.signspeak.eu/), the overall goal is to develop a new vision-based technology for recognizing and translating continuous sign language into text.

Other strategies have focused on recognising signs from the information captured with specialised hardware (Yao, Yao, Liu, Jiang, & August, 2006). However, this is an invasive approach which is sometimes rejected by the Deaf.

In parallel, the Deaf community has found a new communication alternative based on SMS (Short Message Service) languages: not only using mobile phones but also for chat and virtual social networks on the web. The use of SMS (Short Message Service) language was extended with the boom of instant messaging and short message service over the mobile phone. From the communication theory point of view, SMS language is an additional encoding of the message into your own language. Its rapid spread is due to the need to minimize communication costs maintaining the language structure. Generally, deaf people have serious problems with written languages. As SMS languages are simplifications from the written languages, Deaf people have found these short messages easier to understand, finding a communication possibility between hearing and deaf people, especially for young people (Matthews, Young, Parker, & Napier, 2010; Ortiz, 2009)

SMS language is not universal because each language has its own rules in terms of possible abbreviations and phonetics. But, in general, SMS language is characterized by shortening words in relation to the phonetics of the language and their meaning, removing accents and words that are understood by context, deleting "silent" letters such as the 'h' in Spanish, removing punctuation marks, including emoticons, etc.

As a result of expansion of SMS, the need has emerged to develop SMS language to speech translation systems. These systems can be useful in sending SMS messages to fixed phones, with many possible applications. For example, they can be used in emergency situations to send SMSs to people who are not familiar with the SMS language (such as older people), to send messages to visually impaired people or people who are driving or, in this case, to help deaf people to communicate with hearing people. Companies, such as Esendex (http://www.esendex.es/Envio-de-SMS/Voz-SMS) or Comsys (http://www.comsys.net/products-solutions/products/sms-2-fixed.html) currently provide property SMS to speech services, offering the possibility of sending an SMS to a fixed network telephone: the SMS message is translated into speech and sent as a voice message.

Bearing in mind this scenario, this paper describes the development of LSESpeak, 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. The second one is an SMS language to spoken Spanish translation system. This second utility tries to take advance of the SMS communication (widely used by the Deaf for mobile communications) to improve the face to face interaction with hearing people. There are several commercial services for translating SMS language into speech using rule-based strategies. Recently, machine translation strategies are being considered to deal with this problem (Deana, Pennell, & Yang, 2011).

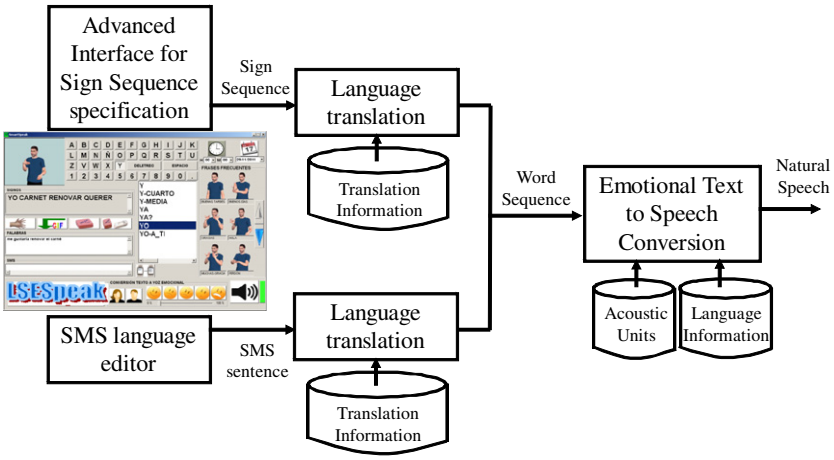

Fig. 1. Module Diagram of LSESpeak.

This paper describes in detail the procedure to develop an open source and free SMS into text translator for Spanish.

When developing LSESpeak, it has been important to keep in mind that Deaf people are a linguistic community without any kind of mental or cognitive problems. Any application for generating spoken language must be differentiated from traditional Augmented and Alternative Communication (AAC) systems for the mentally impaired. Otherwise, Deaf people can feel themselves treated as mentally impaired and reject the developed system. In this sense, the visual interface has to include images to make it simple but it also has to include an important language component, offering all of the flexibility available in their mother tongue (LSE). They have to feel that they are using well-known languages (such as LSE or SMS language), and not a new simplified language especially designed for them.

3. LSESpeak overview

Fig. 1 shows the module diagram of LSESpeak. As it is shown, LSESpeak is made up of two main tools. The first one is a new version of an LSE into Spanish translation system (San-Segundo et al., 2010), 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. 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 (http://www.statmt.org/wmt11/). In the case of the SMS into text translation module, it has been necessary to include pre-processing and post-processing modules in order to deal with some specific characteristics (more details will be described at Section 6).

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). More details about this module will be described in Section 7.

4. Visual interface of LSESpeak

The visual interface design has been the result of a long design process divided into three main steps. In the first step, a group of experts held a brainstorming session to describe the necessary utilities and different alternatives for designing each interface. This group was made up of two experts in LSE (Deaf), a linguist (an expert in Spanish who also knows LSE) and a software developer (who does not know LSE). In the meeting, there were two interpreters for translating Spanish into LSE and vice versa. Secondly, all the alternatives concerning the interface were implemented and evaluated with five end users (Deaf). Every user tested all the design alternatives to generate 15 sentences in LSE using the interface. Finally, there was a meeting including the group of experts and the five users (including the two interpreters). During this meeting, the users were asked about their preferences, drawing up a detailed analysis and the final proposal for the interface.

Fig. 2 shows the main utilities included in the LSESpeak interface.

4.1. LSE and SMS sentence specification

The main utility consists of selecting a letter (clicking on one letter button, i.e., letter Y in Fig. 2) and a list of signs beginning with this letter is displayed in alphabetical order (these buttons have been included by considering the idea of using a touch screen to use the system, instead of the computer keyboard). If a sign from the list is selected (i.e., YO in Fig. 2), the avatar (in the top-left corner) represents it to verify the desired sign corresponding to the sign. In order to add this sign to the sign sequence (the SIGNOS window, under the avatar in Fig. 2), it is necessary to click twice. The sign animation is made using VGuido: the eSIGN 3D avatar developed in the eSIGN project (http://www.sign-lang.uni-hamburg.de/esign/). VGuido has been incorporated as an ActiveX control in the interface.

At any moment, it is possible to carry out necessary actions: to represent the current sign sequence (button), to delete the last sign introduced (button) or to delete the whole sequence (button). Every time the sign sequence is modified, the language translation module is executed and the resulting word sequence is presented in the PALABRAS (words) window. The speak button ( ) executes the TTS converter on the word sequence specified in the PALABRAS (words) window. When the system is speaking, this button is disabled to avoid being used again.

By pressing the DELETREO (spelling) button, the system gets into the spelling mode. In this state, the letter buttons have a different behaviour: they are used to introduce a sign (in the SIGNOS window) letter by letter. This utility is very interesting for specifying a new proper name. When the avatar has to represent a new

Fig. 2. The LSESpeak interface including all of the utilities for LSE or SMS sentence specification and emotional speech generation.

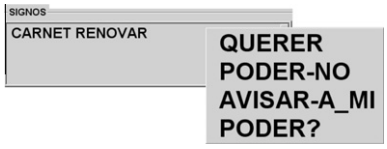

Fig. 3. Example of sign prediction. After including the signs CARNET (LICENCE) and RENOVAR (TO_RENEW), the most probable next signs are QUERER (TO_WANT), PODER-NO (CANNOT), AVISAR-A_MI (REPORT_TO_ME), PODER? (CAN?).

sign, it checks whether there is a text file (in the sign directory) with the sign description. If there are none, the system signs letter by letter. In order to generate several signs by spelling, it is necessary to press the ESPACIO (space) button to separate consecutive signs. In a similar way, it is also possible to specify a number using number buttons and the point button (for decimals).

A very useful utility incorporated into the interface allows the signs following a previous sequence of signs to be proposed. When there is a partial sign sequence specified and the user moves the mouse over the SIGNOS windows, the system displays a popup menu proposing several candidates for the next sign (Fig. 3). These candidates have been selected based on a sign language model trained from sign sequences (LSE sentences) in the corpus. The system proposes the best 4 signs: with highest probability of being selected, given the partial sequence already introduced. If the user clicks on one of these options, the sign is incorporated into the sequence.

In order to facilitate a date or time introduction, the interface offers the possibility of specifying the date on a calendar and the time on a digital clock (Fig. 4). When the date or the time has been specified, it is necessary to push the DATE ( ) or/and TIME ( ) buttons to incorporate the corresponding signs into the sign sequence.

Finally, the visual interface incorporates a list of the most frequent sign sequences (greetings, courtesy formulas, etc.,). When one of these sequences is selected the whole sign sequence is replaced by the selected one, and the SIGNOS and PALABRAS windows are updated (Fig. 5). These sign sequences are represented by animated gifs on the right-hand side of the interface.

This list with the most frequent sign sentences can be extended easily. The visual interface integrates a new utility to generate automatically the gif file necessary to add a new frequent sequence. The gif button ( ) generates a gif file compiling

Fig. 4. Example of date and time introduction. In this example the user has selected the date December, 16th 2011 and the time 12:00.

Fig. 5. Selection of a frequent sign sequence. In this case the selected sentence is BUENOS DÍAS (GOOD MORNING).

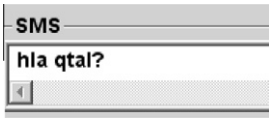

Fig. 6. Example of SMS (hla qtal? "hola ¿qué tal?", in English "hello, how are you?").

the main frames from the representation of the sign sequence specified in the SIGNOS (signs) window.

Finally, in order to specify a SMS sentence, an edition window has been included in the interface (Fig. 6).

Some examples of LSE and SMS input sentences are shown in Fig. 7.

4.2. Emotional text to speech conversion

Fig. 8 shows the main visual controls incorporated to execute the text to speech conversion. The first two buttons allow the voice gender (female or male) to be selected while the following five buttons allow the emotion: happy, sad, angry, surprise and fear to be selected as well. Under these buttons there is a slide control to select the strength of the selected emotion (0% no emotion, 100%

LSE	SMS	Spanish
YO DNI RENOVAR DOCUMENTO NECESITAR CUÁL?	Q ncsito xa rnovar dni?	¿qué necesito para renovar el DNI?
YO CARNET CONDUCIR RENOVAR QUERER DÓNDE?	Dnd rnovar carnt?	¿dónde he de ir para renovar el carné?

Fig. 7. Examples of LSE and SMS inputs and their Spanish translation.

Fig. 8. Visual controls for the emotional text to speech conversion: voice and emotion selection.

highest emotional strength). If neither gender nor emotion is selected, the system will use a neutral male voice.

5. The LSE into Spanish translation module

LSE into Spanish translation is carried out by using a phrase-based translation system based on the software released at the 2011 EMNLP Workshop on Statistical Machine Translation (http://www.statmt.org/wmt11/).

The phrase model has been trained starting from a word alignment computed using GIZA++ (Och & Ney, 2003). GIZA++ is a statistical machine translation toolkit that is used to train IBM Models 1–5 and an HMM word alignment model. In this step, the alignments between the words and the signs in both directions (Spanish-LSE and LSE-Spanish) are calculated: source-target and target-source (LSE-Spanish and Spanish-LSE). Later, a final alignment is generated from a combination of previous alignments. Fig. 9 shows different alignments between a pair of sentences in Spanish and LSE and their alignment points (each black box represents a word and a sign both aligned). The combination can be:

- **Source-Target (ST)**: Only the source-target (LSE-Spanish) alignment is considered. In this configuration, the alignment is guided by signs: each sign in LSE is aligned with a Spanish word and it is possible that some word are unaligned.
- **Target-Source (TS)**: Target-source (Spanish-LSE) is the only considered alignment. In this configuration, the alignment is guided by words: each Spanish word is aligned with a sign in LSE and it is possible that a sign is unaligned.
- **Union (U)**: In this case, the alignment points of the union of both directions (source-target and target-source) are taken. This way, additional alignment points are obtained, creating more examples for training the word translation model, however, the alignment quality is worse (more variability).
- **Intersection (I)**: In this case, the alignment points of the intersection of both directions (source-target and target-source) are selected. This is the strictest configuration: fewer alignment points are obtained, but they are more reliable. This is not a good configuration if there are not enough sentences for training.
- **Grow (G)**: In this configuration, the alignment points of the intersection are used to train the word translation model as well as the adjoining points of union. This configuration is an intermediate solution between the union and intersection, seeking a compromise between the quality and quantity of the alignment points.
- **Diagonal Grow (DG)**: In this configuration, the alignment points of the intersection are considered as well as the adjoining points of the union, but only the diagonal adjoining points.
- **Final Diagonal Grow (FDG)**: In this configuration, the alignment points of the intersection are taken as well as the adjoining points of the union, but only the diagonal adjoining points. And finally, if there is any word or sign unaligned, it is taken from the corresponding union alignment point.

In order to analyse the effect of the alignment in the final results, different alignment configurations were tested in the experiments presented below.

After the word alignment, the system carries out a phrase extraction process (Koehn, 2003) where all of the phrase pairs that are consistent with the word alignment (target-source alignment in our case) are collected. In the phrase extraction, the maximum phrase length has been fixed at seven consecutive words, based on the development experiments carried out on the development set (see the previous section). Finally, the last step is phrase scoring. In this step, the translation probabilities are computed for all phrase pairs. Both translation probabilities are calculated: forward and backward.

The Moses decoder carries out the translation process (Koehn, 2010). This program is a beam search decoder for phrase-based statistical machine translation models. The 3-gram language model has been generated using the SRI language modelling toolkit (Stolcke, 2002).

In order to evaluate the translation module, some experiments have been carried out using the whole Spanish-LSE parallel corpus described in San-Segundo et al. (2010). This corpus contains more than 4,000 parallel sentences (LSE and Spanish) including the most frequent explanations (from government employees) and the most frequent questions (from the user) focused on the domain for renewing an identity document and a driving licence. The corpus was divided randomly into three sets: training (75% of the sentences), development (12.5% of the sentences) and test (12.5% of the sentences) by carrying out a Cross-Validation process. The results presented in this paper are the average of this round robin, increasing the reliability of the results. Table 1 shows the different results for each alignment configuration: mWER (multiple references Word Error Rate), BLEU (BiLingual Evaluation Understudy) and NIST. BLEU and NIST measures have been computed using the NIST tool (mteval.pl).

The best alignment is **target-source**: the alignment is guided by words in this case. The main improvement is due to a fewer number of deletions and these deletions are important in translation because the system translates from a language with fewer tokens per sentence (4.4 in LSE) into a language with more tokens per sentence (5.9 in Spanish). On the other hand, it can be seen that the worst result is given by the intersection alignment, because important alignment points in the target-source are deleted (look at Table 1, most mistakes are deletions). As additional points of target-source are added, the results improve (deletions are reduced), and finally, with target-source the best result is obtained, giving a **3.90% mWER** and a **96.45% BLEU**.

6. The SMS into Spanish translation module

Fig. 10 shows the module diagram of the SMS into Spanish translation system:

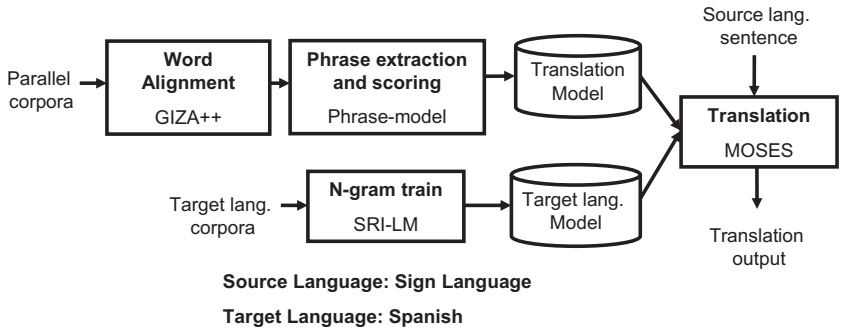

Fig. 9. LSE into Spanish translation architecture.

Table 1
Results of using different alignment configurations.

Alignment	mWER (%)	D (%)	S (%)	I (%)	BLEU (%)	NIST
Intersection	8.41	5.29	1.38	1.75	92.52	11.7069
Sourcer-target	6.52	4.28	1.09	1.14	93.97	11.8033
Diagonal grow	6.39	3.54	1.32	1.53	94.30	11.8022
Union	5.66	2.36	1.96	1.33	94.59	11.7416
Ggrow	5.61	2.34	1.99	1.28	94.59	11.7416
Final diagonal grow	4.84	1.75	2.02	1.07	95.20	11.7218
Target-source	3.90	1.68	1.34	0.89	96.45	11.9020

First of all, there is a pre-processing module that prepares the SMS sentence before sending it to the automatic translator. The second module is the automatic translation system that consists of a phrase-based translator (Moses, the same as that explained in the previous section). The third one is a post-processing module that takes the automatic translator output and deals with the words in this output sentence that have not been translated by the translator, as well as adding the interrogative and exclamatory marks, where necessary.

6.1. Pre-processing and post-processing modules

In the pre-processing module (Fig. 11 left), the first step is to check if there is any question or exclamation mark and, if so, to remove it from the sentence and mark that fact (with the activation of a flag) in order to take it into account in the post-processing. Secondly, the pre-processing check if there is any special character like '+' or '#' next to any term and, if so, the system introduces a space between the character and the term. This action is necessary because, generally, these two isolated characters are translated by the Spanish words "más" (more) and "número" (number), respectively. For example, "q+ kiers?" would be translated into "¿Qué más quieres? (What else do you want?)".

The sentence translated by the phrase-based translator may contain some SMS terms that have not been correctly translated by Moses. Moreover, these sentences do not have question or exclamation marks. Because of this, some post-processing actions are necessary (Fig. 11 right). The first one is to check every term in the translated sentence to see if it is pronounceable or not. This verification is carried out by considering whether a word is pronounceable according to the sequence of consonants and vowels in Spanish (Fig. 12). This verification is carried out by analysing all the sequences of three consecutive letters that make up the term. If one of the three letter sequences is unpronounceable then the term is unpronounceable, otherwise, the term is pronounceable.

If the term is pronounceable, we keep the term in the sentence, because it means that the term has either been translated or has not been translated, but it is a proper name. If the term is not pronounceable, this term is replaced by the Spanish word (selected from a Spanish vocabulary) with the minimum edition distance to the SMS term. The edition distance is calculated with the Levenshtein distance. This distance calculates the minimum

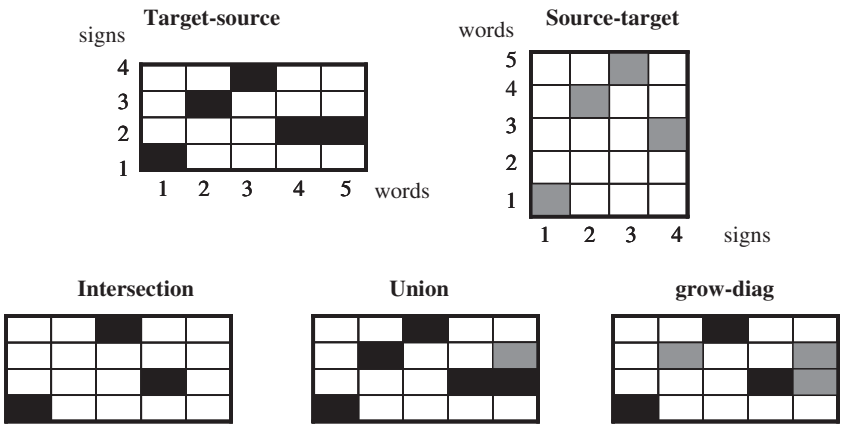

Fig. 10. Different alignment combinations.

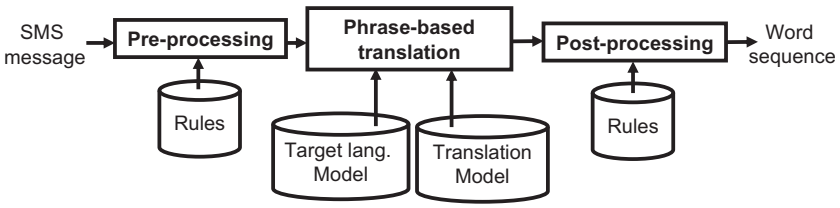

Fig. 11. Diagram of SMS to Spanish translation system.

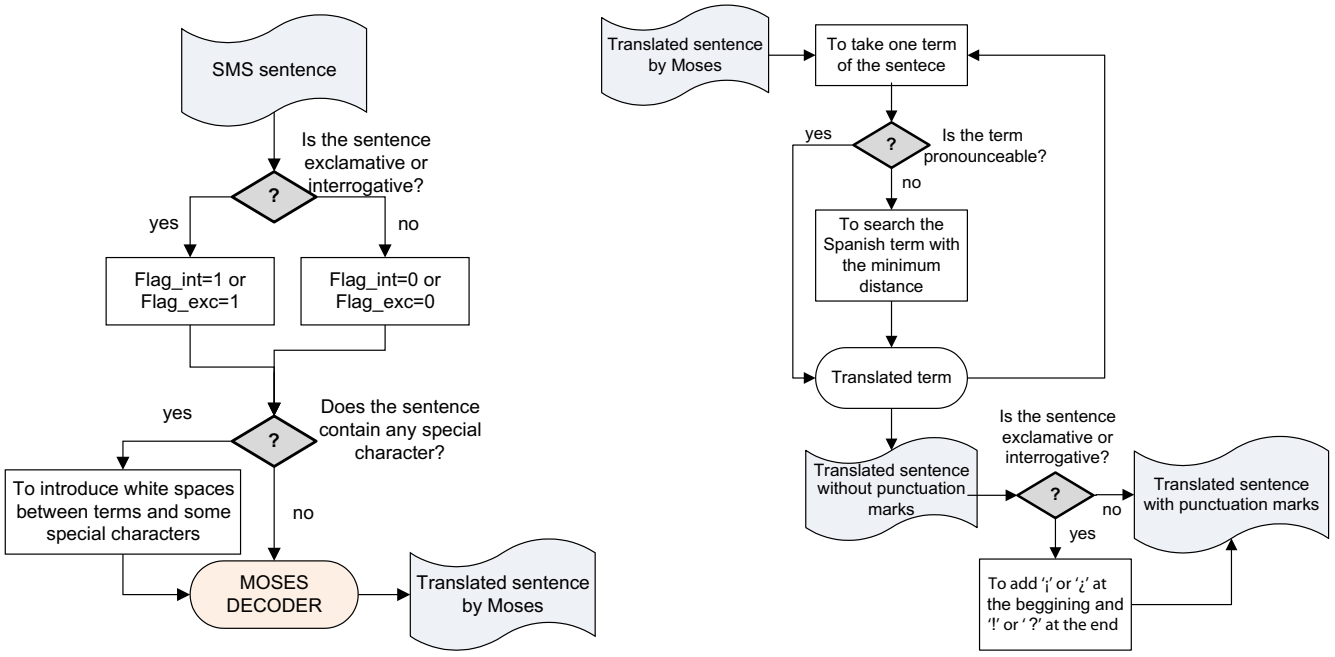

Fig. 12. Diagram of pre-processing (left) and post-processing (right) procedures.

number of operations necessary to transform one string into another: to delete a character, to replace a character with another one, and to insert a new character. Each operation has a cost '1' assigned by default. However, in our system the Levenshtein distance has been modified in order to give less weight to insertions. For this reason, deletions have a cost of 4, substitutions have a cost of 4 and insertions 1, and the final edition distance is the sum of these costs. This is because SMS terms are made up of many deletions (to delete vowels, for example) and substitutions (to substitute some consonants according to the sound of the word) but few insertions of letters. But now, if the translation is carried out on the opposite direction (from SMS language to Spanish), there are many insertions and few deletions. For example, the SMS term "kntndo" has an edition distance to the word "cantando" (singing) of 6: one substitution ('k' for 'c') with cost "4" and two insertions of vowel 'a' with cost "1". The implemented function calculates the edition distance to all of the words in a vocabulary list and it returns the word that has the minimum distance.

Finally, when the translation of the sentence is complete, it is necessary to check whether the sentence is interrogative or exclamatory (indicated by a flag) to add or not question or exclamation marks at the beginning and at the end of the sentence.

6.2. Corpus and experiments

In order to obtain the necessary parallel corpus for training the translation model for the phrase-based translation module, a dictionary of terms extracted from www.diccionariosms.com has been used. This dictionary has been generated by Internet users. This dictionary contains more than 11,000 terms and expressions in SMS language (although this number increases every day) with their Spanish translations and a popularity rate based on the number of users who have registered the term-translation pair. In this way, each SMS word or expression appears with several possible translations. For instance, in Fig. 13 the term "ma" can be translated in Spanish as "madre" (mother), "Madrid", "mama" (mum), "mañana" (tomorrow) or "me ha" (have ... to me). For example, according to popularity, the SMS term "ma" is usually translated into "madre" or "me ha". In this dictionary, there are also emoticons, which are ASCII expressions that mostly represent human faces with certain emotions.

The parallel corpus necessary to train the translation model has been generated from this dictionary. Two files were generated: the first one contains terms and expressions in SMS language and the second one contains Spanish terms that are the corresponding translations. Furthermore, in order to use the popularity as a translation probability measure, each SMS-Spanish pair was weighted in accordance with its popularity number.

In order to prepare this parallel corpus, it was necessary to carry out several actions. The first one was to correct spelling mistakes, checking all of the database carefully. Secondly, SMS sentences can be accompanied by question or exclamation marks. In order to train general models, the proposal was to remove all the question and exclamation marks from the database (except for emoticons).

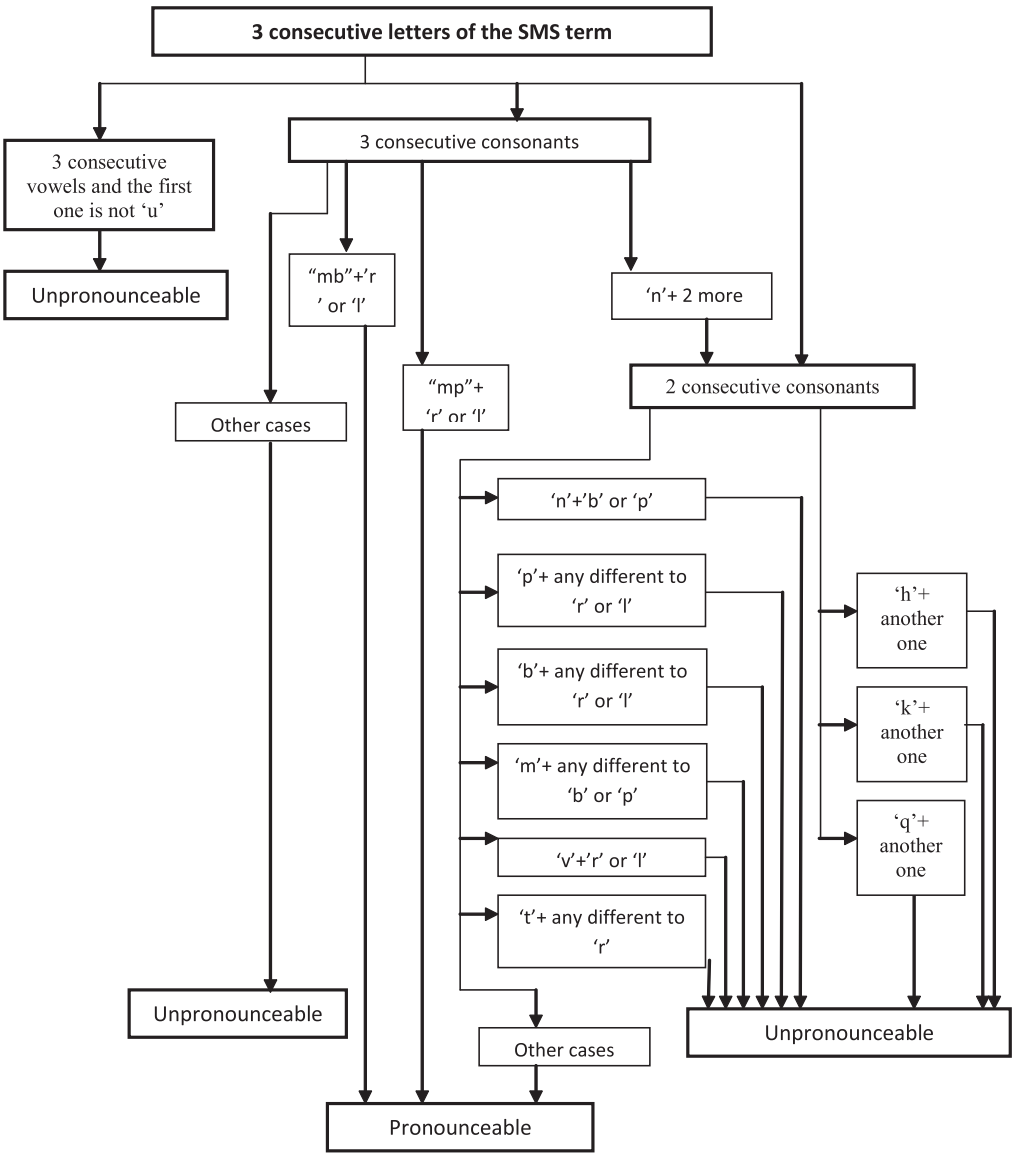

Fig. 13. Function that determines whether one term is pronounceable.

The main idea is to send the sentences to the phrase-based translator removing question and exclamation marks and, later, in a post-process, to add them if it is necessary (see previous section).

Finally, it was necessary to consider that SMS terms (abbreviations of other words) can also be valid words in Spanish. For example, "no" (SMS term for "número" (number) and word for negation in Spanish). Also, when a message is written in SMS language, some words in Spanish are kept without any abbreviation, for instance, short words like "no" or "sí" (yes). For this reason, it is necessary that the training files that contain SMS terms or expressions must also contain Spanish words whose translation is the same word. Therefore, a Spanish vocabulary was incorporated into the parallel corpus (in both files: source and target language files).

In addition to the translation model it is necessary to obtain a target language model in order to select the best translation for each term in its specific context. The first solution was to train the language model considering the Spanish translations (target side) of the parallel corpus (referred to as the "Initial LM"). Secondly, the language model was trained with a general Spanish corpus from the Europarl corpus (http://www.statmt.org/europarl/) (referred to as the "General LM"). In these two situations, the main problem is that, although Deaf people have accepted most of the SMS terms used by hearing people, there are small differences between SMS messages written from Deaf people and SMS written by hearing people. There are differences in the way the words are compressed: typically hearing people tend to remove vowels in a redundant context. For example, "kntndo" is a compression from "cantando" (singing): "ca" is replaced by "k" and "ta" by "t". A deaf person tends to compress verbs using the infinitive: "cantar" (to sing) instead of "cantando". And also, there are differences in the SMS term order: Deaf people sometimes use an order similar to LSE.

In order to adapt the system to these differences, 58 SMS messages from Deaf people in this specific domain (renewal of the Identity Document and Driving License) were collected (referred to as the "Adapted LM"). These messages were translated into Spanish and divided into 8 sub-sets: 7 to adapt the language model and tuning the weights of the models, and the remaining one to evaluate the system. The 8 sub-sets were evaluated with a cross-validation process, calculating the mean of the results. Table 2 shows the different results considering different language models: mWER (multiple references Word Error Rate), BLEU (BiLingual Evaluation Understudy) and NIST.

The final results reveal a significant error reduction when increasing the corpus to train the LM and when adapting it to a specific domain and a specific type of SMS.

7. Emotional text to speech converter

LSESpeak incorporates emotional expression by modelling the emotional speech of the TTS incorporated in the system. The Emotional TTS incorporated based on Hidden Semi-Markov Models provides more flexibility when controlling the Emotional Strength (ES) of the speech output dynamically (according to the needs of the user). This section provides a summary describing the main features of the TTS incorporated in LSESpeak.

7.1. Speech database

The emotional TTS has been developed using the Spanish Expressive Voices (SEV) corpus (Barra-Chicote et al., 2008a). This corpus comprises the speech and video recordings of an actor and an actress speaking in a neutral style and simulating six basic emotions: happiness, sadness, anger, surprise, fear and disgust. SEV presents a relatively large size for a corpus of this type (more than 100 min of speech per emotion). In this work only the speech data of the speaker have been used (almost one hour per emotion). The SEV corpus covers speech data in several genres such as isolated word pronunciations, short and long sentences selected from the SES corpus (Montero et al., 1998), narrative texts, a political speech, short and long interviews, question answering situations, and short dialogues. The texts of all utterances are emotionally neutral. The database has been automatically labeled. Emotional synthetic voices were developed from this corpus using statistical parametric speech synthesis and unit selection synthesis. They have been perceptually evaluated and statistically compared (Barra-Chicote, Yamagishi, King, Montero, & Macias-Guarasa, 2010), achieving an Emotion Identification Rate (EIR) as high as 90%.

7.2. Text processing

In the implementation of the text processing module, the Festival Toolkit has been used. The phone set contains 30 Spanish allophones (including the silence). The modules added to Festival for carrying out the text processing are tokenization and normalization modules (which generate appropriate pronunciations for nouns, acronyms, roman numbers and numbers), a rule-based grapheme to phoneme module (to convert the text input into its phonetic transcription), a rule-based module in charge of automatically splitting the phonetic transcription into a sequence of syllables, a rule-based module for determining whether each syllable from the phonetic transcription is stressed or not, and finally, a categorization module, which discriminates function words from the others.

Using the linguistic processing module, a set of 65 lexical features are extracted at different levels: phoneme level (two previous phonemes, the central phoneme, the two following phonemes, and the position in the syllables of the central phoneme to the following syllable), syllable level (number of phonemes and accent type of the previous syllable, the central syllable and the following syllable; the position of the central syllable in the word and in the phrase; and the identity of the vowel of the syllable), word level (the grammatical category of the previous word, the central word and the following word; number of syllables of the previous word, the central word and the following word; the position of the central word in the phrase (from the beginning and from the end); and the position of the phrase in the sentence), phrase level (number of syllables and words in the previous phrase, the central phrase and the following phrase; and sort of accent in the last phrase), and sentence level (number of syllables, words, and phrases in the sentence).

Table 2
Results considering different language models.

	mWER (%)	BLEU (%)	NIST
Initial LM	46.64	3.82	0.9571
General LM	28.04	35.78	3.8445
Adapted LM	20.18	44.36	4.5331

In order to enhance the basic HSMM-based system, new features coming from a morpho-syntactic analysis of the input sentences have been considered. As the natural language processing (NLP) of the speech synthesis sentences should be very robust (in order to deal with whatever grammatical structures the author of the target texts could use), shallow techniques seem to be a good choice. The first module in our NLP chain is a Spanish Part-Of-Speech tagger (Montero, 2003), based on ESPRIT-860's EAGLES-like 10-byte labels (more than 250 possible tags), using a set of dictionaries such as the Spanish-Academy (RAE) 159,898-word dictionary, the richly-tagged ESPRIT-860 9883-word dictionary, Onomastica's 58,129-proper-noun dictionary, GTH's 1960-multiword expression dictionary, and GTH's verb conjugation analyzer (including 102 irregular paradigms and 4325 infinitives).

After assigning all possible tags to the words, several sets of hand-written rules are used for cascade-filtering impossible tag sequences: GTH's 77 high-recall rules, CRATER's 148 rules and GTH's 162 low-recall high-precision rules. On the 38,172 word test-set of the ESPRIT-860 Spanish corpus, the recall is as high as 0.9987 when averaging 1.6984 tags per word.

Finally, the Trigrams N Tags (TnT) stochastic tagger is used for disambiguation. This tagger uses an interpolated language model based on trigrams, bigrams and unigrams, resulting in a 98.99% accuracy for a 1-tag-per-word basis, or 99.45% if 1.0238 tags are assigned per word on average. After tagging the sentence, 2 features are available to be used in the speech synthesis training and testing process: a gross 10 category feature (based on an 860 set of tags) and a 3-byte tag from the 860 coding scheme (including a first byte for the 10 main tags and 2 additional bytes for a more detailed subtagging).

The final NLP processing module is a shallow parser based on a Cocke-Younger-Kasami (CYK) bottom-up algorithm and a set of 2179 hand-written general-purpose CYK parsing rules. As these rules are very ambiguous, many possible parser trees are assigned to each sentence. In order to control the exponential growth of this analysis, a small set of splitting rules were developed (trying to reduce the length of the text to be analyzed) and a final filtering process was used, selecting only one tree using a Minimum Description Length approach. In a subset of the test set, for a total 5703 shallow syntactic phrases, there were 0.35% cutting errors, 0.55% tagging-recall errors, 1.10% tagging-precision errors and 1.49% syntactic-analysis errors. These shallow syntactic phrases are the third feature to be used in the synthesis process.

Fig. 14 shows the results of a perceptual test to validate the quality improvements when adding the morpho-syntactic features. The results reveal a 41% statistically significant relative improvement in the Mean Opinion Score (MOS).

7.3. Acoustic modelling

The HSMM-based voices were built using a system widely described in Barra-Chicote et al. (2008b). The architecture of the system is detailed in Fig. 15. The HSMM-based speech synthesis system comprises three components. The first one is the speech

SMS term	Spanish translations	Popularity
ma	madre (mother)	2
	Madrid (Madrid)	1
	mamá (mum)	8
	mañana (tomorrow)	1
	me ha (have .... to me)	8

Fig. 14. Segment of the SMS-Spanish web dictionary.

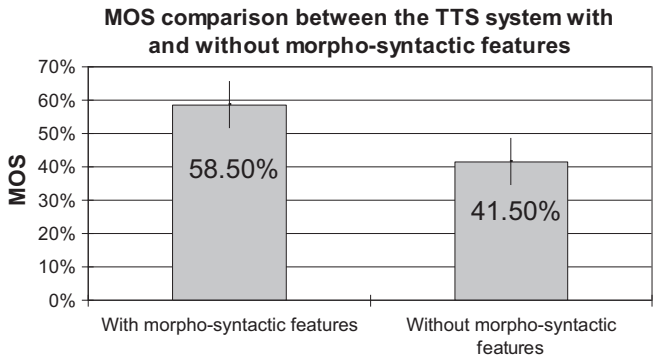

Fig. 15. MOS comparison between the TTS system with and without morpho-syntactic features, respectively.

analysis part, where three kinds of parameters for the STRAIGHT mel-cepstral vocoder with mixed excitation (the mel-cepstral coefficients, log F0 and a set of aperiodicity measures) are extracted as feature vectors to be modelled by the HSMMs. These are as described in Zen, Toda, Nakamura and Tokuda (2007), except that the F0 values used were more robustly estimated using a vote amongst several F0 extraction algorithms (Yamagishi et al., 2009). Secondly, the HSMM training part, where context-dependent multi-stream left-to-right MSD-HSMMs are trained for each emotion using the maximum likelihood criterion. In the speech generation part, acoustic feature parameters are generated from the MSD-HSMMs using the GV parameter generation algorithm. Finally, the speech generation part, where an excitation signal is generated using mixed excitation (pulse plus band-filtered noise components) and pitch-synchronous overlap and add (PSOLA). This signal is used to excite a mel-logarithmic spectrum approximation (MLSA) filter corresponding to the STRAIGHT mel-cepstral coefficients, generating the speech waveform.

7.4. Dynamic emotional strength manipulation

Emotional independent models can be used straightforwardly to synthesize an utterance with emotional content. However, since state-of-the-art systems focus on corpus-based synthesis techniques (this means that the whole knowledge is learnt from data; instead of using rule-based strategies based on human expertise) and emotional models are usually built from acted data. Using those models, the speech output tends to be perceived as emotionally exaggerated speech. In order to avoid this barrier to naturalness, mainly due to the acted material (real emotional data is extremely difficult to achieve), LSESpeak incorporates a slider to control the emotional strength dynamically. The explicit modeling of speech used by the HSMM-based synthesis, allows the emotional strength of the speech output to be modified by using model interpolation techniques (Yamagishi, Onishi, Masuko, & Kobayashi, 2005), producing an enhanced natural expressive response.

Fig. 16 shows the process in charge of controlling and applying the selected emotional strength. The emotional strength selected by the user by means of the slide control available on the graphical interface is filtered in an attempt to avoid strength over-smoothing and over-emphasis. A set of linguistic labels are extracted by applying the linguistic processing modules to the word sequence obtained as the output of the Sign to Text module. The emotional model selected by the user on the graphical interface is interpolated with the speaker's neutral voice by means of a weighted combination of both models.

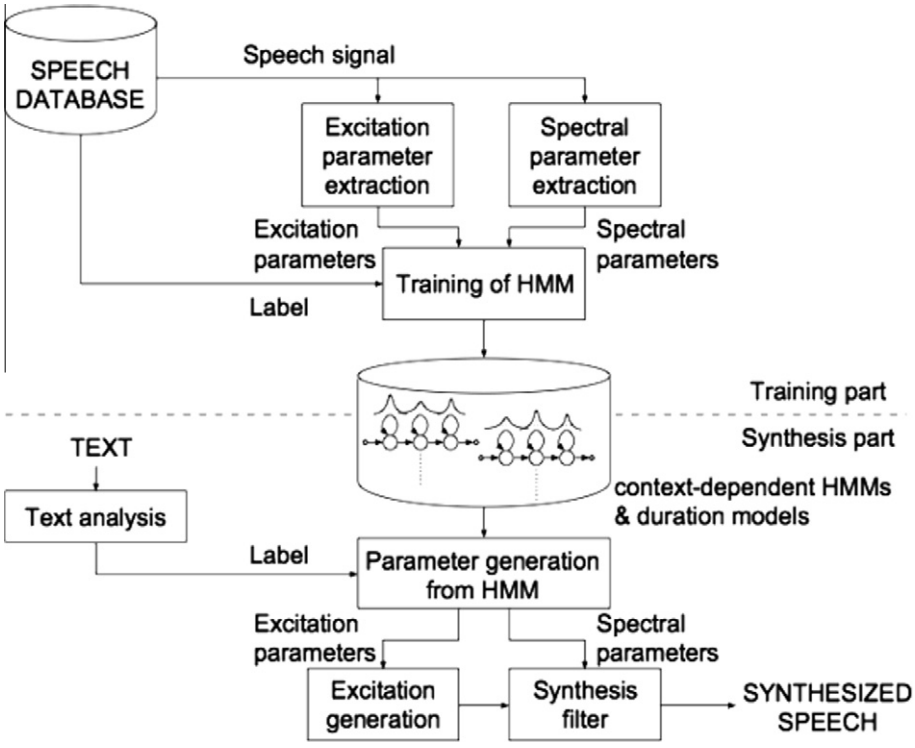

Fig. 16. Block diagram of the text to speech synthesis.

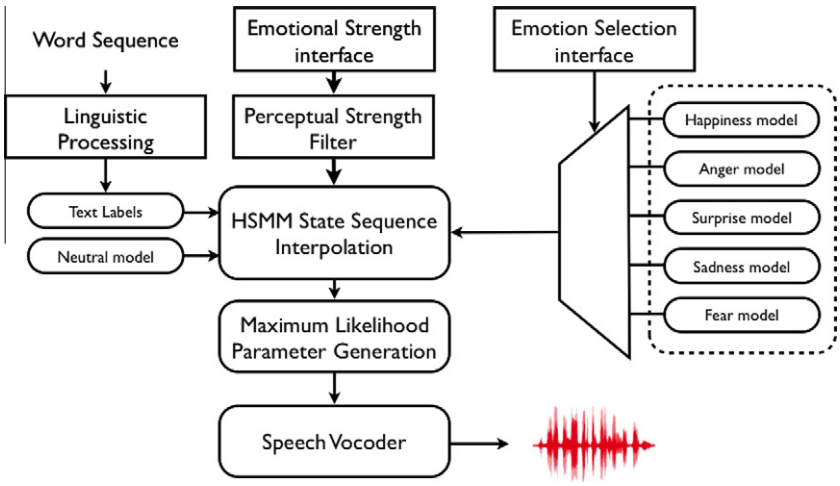

Fig. 17. Emotional text to speech synthesis process with dynamic emotional strength manipulation.

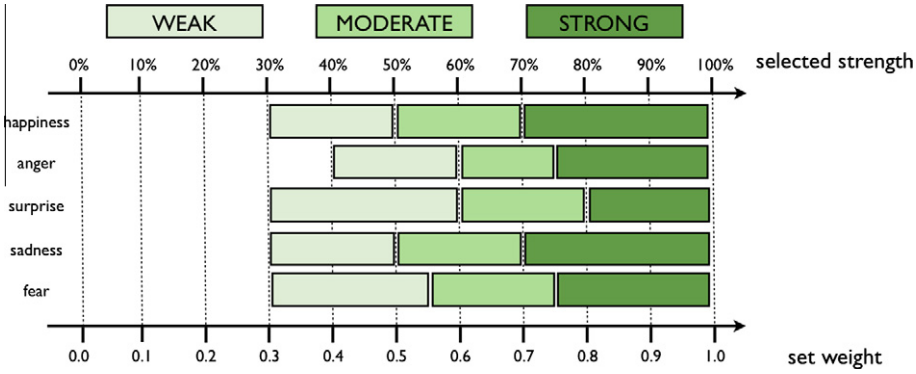

Fig. 18. Perceptual results of the categorization of the perceived emotional strength.

The models' interpolation module uses the linguistic labels to generate the optimal HSMM state model sequences using both models independently, neutral and emotional model respectively. Then, the two state sequences, are interpolated into an output sequence of Probability-Density-Functions (PDFs), represented by a mean vector $\hat{\mu}$ and covariance matrix $\hat{\sum}$:

$$\hat{\mu} = \sum_{k=1}^{N} a_k \mu_k, \hat{\Sigma} = \sum_{k=1}^{N} a_k^2 \Sigma_k \quad (1)$$

where μ_k and $\sum_k$ are the mean vector and covariance matrix of the output sequence PDFs of model k, and a_k is the weight applied to each model as established by the perceptual strength filter module.

The target emotional strength, selected by the user by means of the slide control provided on the graphical interface (Fig. 7), should be filtered by attending to the perception of the target strength from the user system. In order to learn the relationship between the interpolation weight and the perceived emotional strength, a perceptual test was carried out. Twenty listeners, having a similar socio-cultural profile, participated in the evaluation, which was carried out individually in a single session per listener. All listeners, between twenty and thirty years old, were from the Madrid area, and none of them had a speech-related research background nor had they previously heard any of the SEV speech recordings.

The test consisted of the evaluation of the perceived emotional strength of three different utterances for every emotion. Listeners could synthesize each utterance with a specific interpolation weight (the interpolation weight is the selected emotional strength, normalized between 0 and 1) as many times as deemed necessary. For every emotion, listeners were asked to define the boundaries of three strength categories: week, moderate and strong. The resulting boundaries for each strength category and emotion are plotted in Fig. 17.

As can be seen in the figure, most of the emotions are not perceived when selecting strength levels below 30%, except for anger, which is not perceived below 40%. The emotional strength of all emotions is moderately perceived when applying interpolation weights between 0.4 and 0.5, and starts to be strongly perceived when using interpolation weights higher than 0.8. Emotional strength of anger and surprise is perceived as smoother than the other emotions: moderate strength starts at a 60% level (instead of 50%); and the lower boundaries for strong starts at 75%. This is especially interesting, since the low strength level of surprise makes this emotion become intrinsically confused with happiness (both are positive emotions) (see Fig. 18).

These perceptual results suggest that it would interesting to incorporate these emotional strength categories into the user interface, in order to allow the emotional strength level to be controlled by applying, for example, the average interpolation-weight of each interval obtained in this evaluation for each emotion.

8. Conclusions

This paper has described LSESpeak, a new application for helping Deaf people to generate spoken Spanish. This application

includes two main utilities: the first one is an advanced version of a spoken Spanish generator from LSE. This tool includes a visual interface where Deaf people can specify an LSE sentence. This sentence is translated into Spanish and then passed to a TTS to generate spoken Spanish. In this new version, new utilities have been included: removing useless utilities, incorporating icon based buttons, representing frequent sentences with animated gifs, automatic generation of gifs for new sentences and a better language translation module. As regards the translation module, this work has investigated the effect on the performance of the sign-word alignment strategy to train the translation model. The best alignment configuration has been target-source where alignment is guided by words: fewer deletions are produced in translated sentences. It can also be seen that when the target-source alignment points are removed, the word error rate increases. The cause is the increase in word deletions at the output. These deletions are important when translating from LSE (with fewer tokens per sentence) to Spanish (with more tokens per sentence). For the best configuration, the system obtains a 3.90% mWER and a 96.45 BLEU.

The second utility incorporated into LSESpeak is an SMS language to spoken Spanish translation system. This tool is made up of an editing window for writing the SMS, a language translation module and an emotional text to speech converter. The language translator uses a phrase-based translation strategy similar to the previous tool, but in this case, it has been necessary to include pre-processing and post-processing modules for dealing with specific characteristics: text normalization and dealing with unpronounceable terms. Experiments presented in Section 6.2 reveals a significant error reduction when the system is adapted to a specific domain and a specific type of SMS.

Related to the emotional text to speech converter, it is important to highlight the increase in naturalness obtained when incorporating morpho-syntactic features, and the high flexibility provided by HSMMs when generating different emotional strengths.

Acknowledgments

The work leading to these results has received funding from the European Union under grant agreement n 287678. It has also been supported by TIMPANO (TIN2011-28169-C05-03), ITALIHA (CAM-UPM), INAPRA (MICINN, DPI2010-21247-C02-02), SD-TEAM (MEC, TIN2008-06856-C05-03) and MA2VICMR (Comunidad Autónoma de Madrid, S2009/TIC-1542) projects. Authors also thank Mark Hallett for the English revision of this paper and all the other members of the Speech Technology Group for the continuous and fruitful discussion on these topics.

References

Barra-Chicote, B., Montero, J. M., Macias-Guarasa, J., Lufti, S., Lucas, J. M., Fernandez, F., D'haro, L. F., San-Segundo, R., Ferreiros, J., Cordoba, R., & Pardo, J. M. (2008a). Spanish expressive voices: Corpus for emotion research in Spanish. In *Proceedings of 6th international conference on, Language Resources and Evaluation*, May 2008.

Barra-Chicote, R., Yamagishi, J., Montero, J. M., King, S., Lufti, S., & Macias-Guarasa, J. (2008b). Generacion de una voz sintetica en Castellano basada en HSMM para la Evaluacion Albayzin 2008: conversion texto a voz. In *Proceedings of V. Jornadas en Tecnologia del Habla, November 2008. (Best system award)*.

Barra-Chicote, R., Yamagishi, J., King, S., Montero, J. M., & Macias-Guarasa, J. (2010). Analysis of statistical parametric and unit selection speech synthesis systems applied to emotional speech. *Speech Communication, 52*(5), 394–404.

Deana, L., Pennell & Yang, Liu. (2011). A character-level machine translation approach for normalization of SMS abbreviations. *IJCNLP*.

Dreuw, P. (2008). Visual modeling and tracking adaptation for automatic sign language recognition. In *Proceedings of international computer vision summer school (ICVSS)*. Sicily, Italy, July 2008.

Dreuw, P., Forster, J., Deselaers, T., & Ney, H. (2008). Efficient approximations to model-based joint tracking and recognition of continuous sign language. In *Proceedings of IEEE International Conference Automatic Face and Gesture Recognition (FG)*. Amsterdam, The Netherlands, September 2008.

Dreuw, P., Stein, D., & Ney, H. (2009). Enhancing a sign lLanguage translation system with vision-based features. In *Proceedings of special issue gesture workshop 2007, LNAI*. (Vol. 5085, pp. 108–113). Lisbon, Portugal, January 2009.

Dreuw, P., Ney, H., Martinez, G., Crasborn, O., Piater, J., Miguel Moya, J., & Wheatley, M. (2010). The sign-speak project - bridging the gap between signers and speakers. In *Proceedings of 4th workshop on the representation and processing of sign languages: Corpora and sign language technologies (CSLT 2010)* (pp 73–80). Valletta, Malta.

Dreuw, P., Forster, J., Gweth, Y., Stein, D., Ney, H., Mar-tinez, G., Verges Llahi, J., Crasborn, O., Ormel, E., Du, W., Hoyoux, T., Piater, J., Moya Lazaro, J. M., & Wheatley, M. (2010). SignSpeak - Understanding, recognition, and translation of sign languages. In *Proceedings of 4th workshop on the representation and processing of sign languages: Corpora and sign language technologies (CSLT 2010)* (pp 65–73). Valletta, Malta, May 2010b.

Efthimiou, E., Fotinea, S., Hanke, T., Glauert, J., Bowden, R., Braffort, A., Collet, C., Maragos, P., & Goudenove, F. (2010). DICTA-SIGN: Sign language recognition, generation and modelling with application in Deaf communication. In *Proceedings of 4th workshop on the representation and processing of sign languages: corpora and sign language technologies (CSLT 2010)* (pp 80–84). Valletta, Malta, May 2010.

Hanke, T., König, L., Wagner, S., Matthes, S. (2010). DGS Corpus & Dicta-Sign: The hamburg studio setup. In *Proceedings of 4th workshop on the representation and processing of sign languages: corpora and sign language technologies (CSLT 2010)* (pp. 106–110). Valletta, Malta, May 2010.

Karami, A., Zanj, B., & Sarkaleh, A. (2011). Persian sign language (PSL) recognition using wavelet transform and neural networks. *Expert Systems with Applications, 38*(3), 2661–2667.

Koehn ,P., Och, F.J., & Marcu, D. (2003). Statistical phrase-based translation. In *Proceedings of human language technology conference 2003 (HLT-NAACL 2003)* (pp. 127–133) Edmonton, Canada.

Koehn, P. (2010). Statistical machine translation. Ph.D. thesis. Cambridge University Press.

Matthews, N., Young, S., Parker, D., & Napier, J. (2010). Looking across the hearing line?: Exploring young Deaf people's use of Web 2.0. *M/C Journal, 13*(3).

Montero, J. M., Gutierrez-Arriola, J. M., Palazuelos, S., Enriquez, E., Aguilera, S., & Pardo, J. M. (1998). Emotional speech synthesis: From speech database to TTS. In *Proceedings of ICSLP-98* (pp. 923–926). December 1998.

Montero, J. M. (2003). Estrategias para la mejora de la naturalidad y la incorporación de variedad emocional a la conversión texto a voz en castellano. Ph.D. thesis, Escuela Técnica Superior de Ingenieros de Telecomunicación, Speech Technology Group,Electrical Engineering Department at Universidad Politecnica de Madrid, Spain, May 2003.

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, 32*(1), 24–37.

Och, J., & Ney, H. (2003). A systematic comparison of various alignment models. *Computational Linguistics, 29*(1), 19–51.

Ortiz, T. (2009). Local turns global: Expanding the Deaf community through communication technologies. *In Proceedings of TCC 2009*.

San-Segundo, R., Pardo, J. M., Ferreiros, F., Sama, V., Barra-Chicote, R., Lucas, JM., et al. (2010). Spoken Spanish generation from sign language. *Interacting with Computers, 22*(2), 123–139.

San-Segundo, R., Montero, J. M., Córdoba, R., Sama, V., Fernández, F., D'Haro, L. F., et al. (2011). Design, development and field evaluation of a Spanish into sign language translation system. *Pattern Analysis and Applications, 15*(2), 203–224. 2012.

Stolcke, A. (2002). SRILM – An extensible language modelling toolkit. In *Proceedings of International Conference on spoken language processing* (Vol. 2, pp. 901–904). Denver, USA.

Sylvie, O., & Surendra, R. (2005). Sign language analysis: A survey and the future beyond lexical meaning. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 27*(6), 12.

Tryggvason, J. (2004). VANESSA: A system for council information centre assistants to communicate using sign language. School of Computing Science, University of East Anglia.

Vogler, C., & Metaxas, D. (2001). A framework for recognizing the simultaneous aspects of ASL. *CVIU, 81*(3), 358–384.

von Agris, U., Schneider, D., Zieren, J., & Kraiss, K.-F. (2006). Rapid signer adaptation for isolated sign language recognition. In *Proceedings of CVPR workshop V4HCI* (pp. 159). New York, USA, June 2006.

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

Wheatley, M., & Pabsch, A. (2010). Sign Language in Europe. In *4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies*. W. Stokoe, Sign Language structure: an outline of the visual communication systems of the American deaf. Studies in Linguistics, Buffalo University. Paper 8, 1960. LREC Malta 2010.

Yamagishi, J., Onishi, K., Masuko, T., & Kobayashi, T. (2005). Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis. *IEICE Transactions on Informations and & Systems, E88-D*(3), 503–509.

Yamagishi, J., Nose, T., Zen, H., Ling, Z-H., Toda, T., & Tokuda, K. (2009). A robust speaker-adaptive HMM-based text-to-speech synthesis. *IEEE Trans. Speech, Audio & Language Process*, 2009.

Yao, G., Yao, H., Liu, X., & Jiang, F. (2006). Real time Large vocabulary continuous sign language recognition based on OP/Viterbi Algorithm. In *Proceedings of 18th ICPR, August 2006* (Vol. 3, pp. 312–315).

Yung-Hui, Lee, & Cheng-Yue, Tsai (2009). Taiwan sign language (TSL) recognition based on 3D data and neural networks. *Expert Systems with Applications, 36*(2), 1123–1128. Part 1.

Zen, H., Toda, T., Nakamura, M., & Tokuda, K. (2007). Details of Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005. In *Proceedings of IEICE Transactions on Informations & Systems* (Vol. E90-D(1), pp. 325–333). January 2007.