Integration of a Talking Head into a Spanish Sign Language Synthesizer

Javier Tejedor, Fernando López-Colino, Javier Garrido and José Colás

Human Computer Technology Laboratory, Escuela Politécnica Superior,
Universidad Autónoma de Madrid, Spain
[javier.tejedor, fj.lopez, javier.garrido, jose.colas]@uam.es

Abstract. In this paper, we present an integration of a talking head within a Spanish Sign Language synthesizer. The whole system consists of three different steps: First, the input acoustic signal is transformed into a sequence of phones by means of a speech recognition process. This sequence of phones is mapped in a second step to a sequence of visemes and finally, the resulting sequence of visemes is played by means of a talking head integrated into the avatar used in the Spanish Sign Language synthesizer.

Key words: Spanish Talking Head, Speech Recognition, Visemes, Spanish Sign Language

1 Introduction

Hard of hearing people cannot access to the information in an identical fashion as those who do not suffer from such disability. Several efforts have been done to solve this issue by means of a talking head which represents what a person is speaking. Some of them are presented in [1–5]. Others are in an early development stage [6]. However, the talking head approach is not enough for deaf people, as lip reading is used as a complement to Sign Language communication. Despite of that, minimum efforts have been done for Spanish language by means of systems that convert the speech-based content into a sequence of signs to be represented by a Spanish Sign Language (SpSL) synthesizer [7, 8], while others are in an early development stage [9, 10].

In this paper, we propose to augment the functionality of our SpSL synthesizer, which receives its input using a specifically designed XML-formatted message, by adding the required information to emulate lip movements to allow the deaf people to access to the speech-based content. The talking head of the signing avatar produces the lip movement required according to the output of the recognition process which represents the information stored in the speech signal.

The rest of the paper is divided as follows: Section 2 presents an overview of the system. Section 3 presents the work for speech recognition. Section 4 presents the mapping between phone and viseme. Section 5 describes the SpSL synthesizer. Section 6 presents the experimental setup. Section 7 presents the evaluation and finally Section 8 presents our conclusions.

2 System Overview

The whole system architecture is depicted in Fig. 1. It consists of three different modules: i) The *Speech Recognition* module converts the input acoustic signal into the most likely sequence of phones identifying it. ii) The *Phone-to-Viseme Mapping* module takes the sequence of phones hypothesized by the *Speech Recognition* module and maps it to the sequence of visemes. iii) The talking head, integrated into the *SpSL Synthesizer* module, takes this sequence of visemes and produces the lip movement according to it within the image that represents the avatar used in the SpSL synthesizer. The next sections describe each module in more detail.

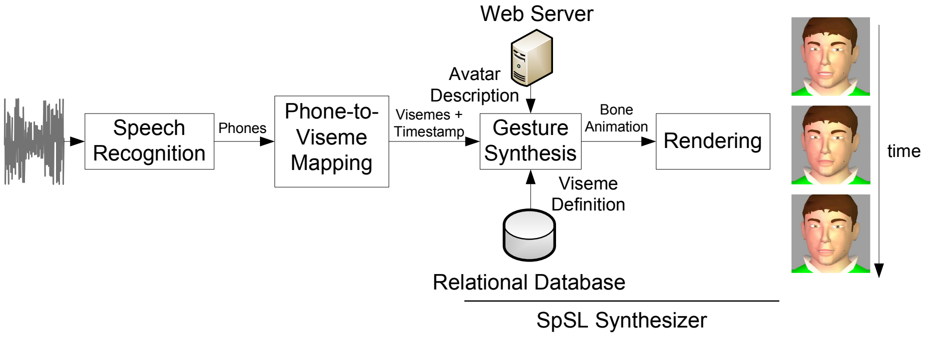


Fig. 1. Schematic view of the system architecture.

3 Speech Recognition

The aim of the speech recognition is to produce the most likely sequence of phones corresponding to the input acoustic signal. The HTK tool [11] was used for the speech recognition work. The Viterbi algorithm takes the input acoustic signal, the set of phones to recognize, the Hidden Markov Models (HMMs) used to represent the acoustic space, the language model composed of phones and hypothesizes the sequence of phones according to the input acoustic signal. This sequence is passed towards the *Phone-to-Viseme Mapping* module.

4 Phone-to-Viseme Mapping

To build the sequence of visemes to be represented by the talking head, a phone-to-viseme mapping is necessary. There exists a regular mapping between the set of phonemes and the set of visemes in Spanish, which is represented in Table 1. An additional neutral viseme was added to represent the silence of the speech signal and the initial position of the lips in the talking head.

First, the set of phones output by the *Speech Recognition* module was transformed to each corresponding phoneme in the Table 1. After that, such sequence of phonemes serves to produce the sequence of visemes according to the mapping specified in the Table 1 as output. This sequence of visemes is passed towards the *SpSL Synthesizer* module. Fig. 2 depicts both the speech recognition and the phone-to-viseme mapping processes.

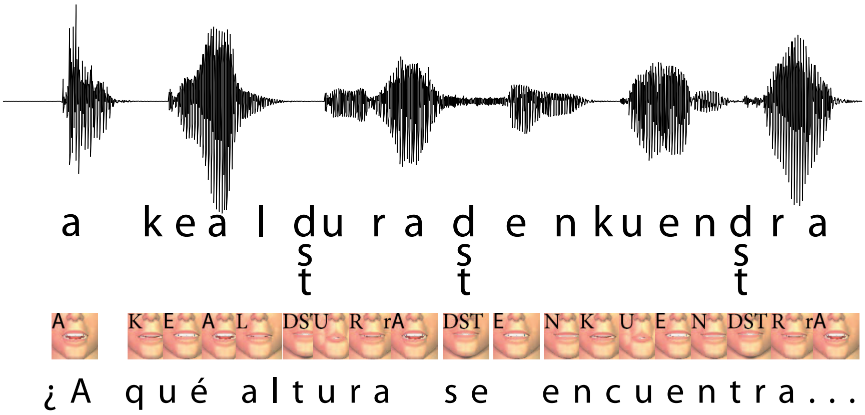


Fig. 2. Speech signal, the sequence of visemes mapped from the phones recognized and the actual ortographic transcription of the speech signal.

5 SpSL Synthesizer

The structure of the SpSL signing avatar was described in [9] as a hierarchical bone structure whose movements define the deformation of the avatar's mesh. This bone structure has been also used in the definition of the avatar's face and mouth. This approach has some advantages against the animation based on the interpolation of morphing objectives. The first advantage is related to storage requirements. Each morphing objective is a copy of the avatar's mesh in which some vertexes have been displaced to represent a different shape. The bone animation approach requires defining the avatar's mesh just once and storing the orientations of the skeleton bones, which causes a reduction in the storage requirements. The second advantage of this approach refers to the addition or modification of the mouth shapes. The morphing approach requires the release of a new version of the avatar including the new meshes. The definition of a new mouth shape using the bone animation approach only requires defining new orientations for the relevant bones for the mouth shape definition. Finally, the last advantage in using this bone animation approach in the signing avatar relies on the unification of the animation techniques applied to the avatar. This simplifies the development of the SpSL synthesizer.

Viseme	Phoneme
a	a
e	e
i	i
o	o
u	u
f	f
g	ɡ
j	x
l	l
bmp	b, m, p
chy	ʧ, ʎ, ɟ, x
dst	d, s, t, θ
k	k
n	n, ɲ
r	ɾ, r

Table 1. Relationship between the set of phonemes and the set of visemes in Spanish.

5.1 Avatar's Mouth Description

We propose a structure of seven bones to define the shape of the mouth. The only anatomic bone implied in the definition of the mouth shape is the jaw because the flexion of the Temporomandibular joint defines the openness of the mouth. The first of the seven bones used for this task emulates the jaw, meanwhile the other six bones emulate variations of the mouth shape defined by the contractions of the muscles surrounding the mouth.

Using the Fig. 3 we can describe the location and function of the six bones that emulate the contraction of the lip-related muscles. Upper and lower lips require two different bones each: the first bone affects a wider area because it simulates the retraction movement of the lip (areas 1 and 5 of the mouth). The second bone is used for the elevation movement of the bones (e.g. saying "u"), which affects a narrower area of the upper and lower lips (areas 2 and 4). The last two bones, areas 3 and 6, modify the position of the commissures without affecting the middle part of the mouth.

The tongue is animated using three different bones: the first one establishes the position of the mouth and the other two bones define the flexion of the tongue muscles. Currently we do not generate any animation for the tongue because these movements are hardly seen.

5.2 Gesture Synthesis

Each viseme requires defining the orientation of all the mouth shape-related bones. This information is stored in the same relational database that stores SpSL sign definitions. Several signs require a specific mouth shape as part of their definition. These mouth shapes are considered to be a part of the Non

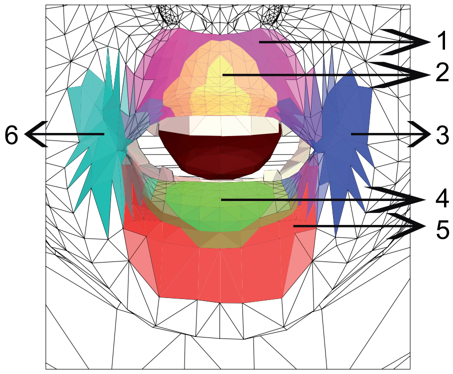


Fig. 3. This figure shows the different mouth areas affected by the different mouth bones.

Hand sign parameter. Therefore, the generation of the lip movements to emulate speech is included within the whole sign generation process easily.

The output of the *Phone-to-Viseme Mapping* module is a sequence of visemes with a timestamp. This sequence is processed in the *Gesture Synthesis* module as a part of the whole SpSL synthesis process. For each viseme in the sequence, one animation keyframe is defined for each bone of the mouth. The timestamps obtained from the viseme sequence are directly assigned to the corresponding keyframe. Grouping all the keyframes belonging to a bone defines the animation track of that bone. The seven animation tracks related to the mouth bones (referred as Bone animation in Fig. 1) are required in the *Rendering* module for the generation of the final animated visualization.

5.3 Rendering

The 3D rendering module is based on the JSR-184 standard 3D API definition [12]. This module generates the final visualization using the description of the avatar and the animation tracks created in the *Gesture Synthesis* module. This visualization consists of a sequence of static images. If the animation instant of one of these images does not coincide with a keyframe, the orientation of the bones is defined using interpolation techniques. Fig. 4 shows our standard avatar visualization representing the sign TÚ (you) while the mouth shape represents the viseme “U”. This visualization uses a far camera shot.

The approach used to interpolate the keyframes is the linear interpolation, where the orientation of every bone between the keyframes i and $i+1$ is defined linearly using the values of these two keyframes.

6 Experimental Setup

6.1 Speech Recognition

The input acoustic signal is sampled at 16kHz and stored with 16 bit precision. Mel Frequency Cepstral Coefficients (MFCCs) were computed at 10ms intervals within 25ms Hamming windows. Energy and first and second order derivatives were appended giving the 39 MFCCs used to represent the signal.

The set of 47 phones in Spanish language [13] has been used during the recognition process. HMMs were used as acoustic models to represent the set of phones and they were context-dependent with 8-components Gaussian Mixture Models (GMMs).

The Albayzin database [14], which contains two different sub-corpora, has been used in the experiments: a phonetic corpus and a geographic corpus. Each contains a training set and a test set. The training of the HMMs was made from the ***phonetic training set***. A bigram was used as language model during the phone recognition process. It was built from the ***phonetic training set*** as well. The number of components GMMs in the context-dependent acoustic models was tuned for phone accuracy on the ***phonetic test set***. The parameters *word insertion penalty* and *language scale factor*, used within the recognition process, were tuned on the ***geographic training set***. Finally, the ***geographic test set*** was used for the system evaluation. Fifteen sentences, which contain between 4 and 17 words each, corresponding to 15 audio files in the ***geographic test set*** were used in the whole system evaluation.

6.2 Spanish Sign Language Synthesizer Talking Head

A score of the naturalness of the talking head in the range $[1, 5]$ has been used in the evaluation for the talking head when the audio file is also presented to the listener. We have selected two different camera shots for two different evaluations. The one referred as *Near camera* in Table 2 uses the picture in Fig. 5(a). The second one is referred as *Far camera* in Table 2 and is depicted in Fig. 5(b). Both evaluations were made with 30 Spanish listeners whose ages vary from 18 to 57 years old.

7 Evaluation

The accuracy of the speech recognition module plays a very important role in the final system performance. A mis-recognized phone leads to an incorrect talking head lip position, which affects the final system performance. The phone recognition accuracy was 70.61% in the set of 15 audio files chosen for the evaluations, which led to a 83.02% of accuracy in the set of visemes. It means that at about 1 out of 5 visemes will cause an error in the sequence of visemes to be represented by the avatar.

The results presented in Table 2 show an improvement in using the near camera. Paired t-tests showed that such difference is statistically significant with

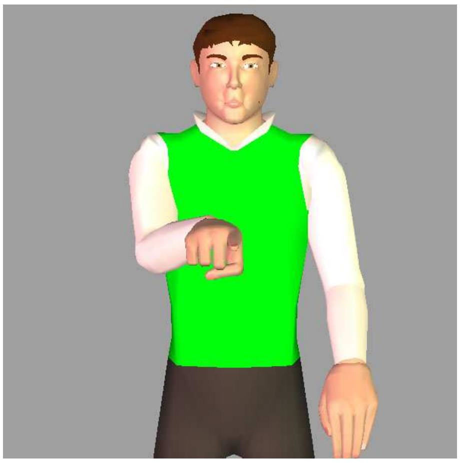


Fig. 4. The SpSL avatar includes the possibility of lip movement emulation.

	Near camera	Far camera
Score	3.66	3.53

Table 2. Results for the Spanish talking head evaluation from the two cameras.

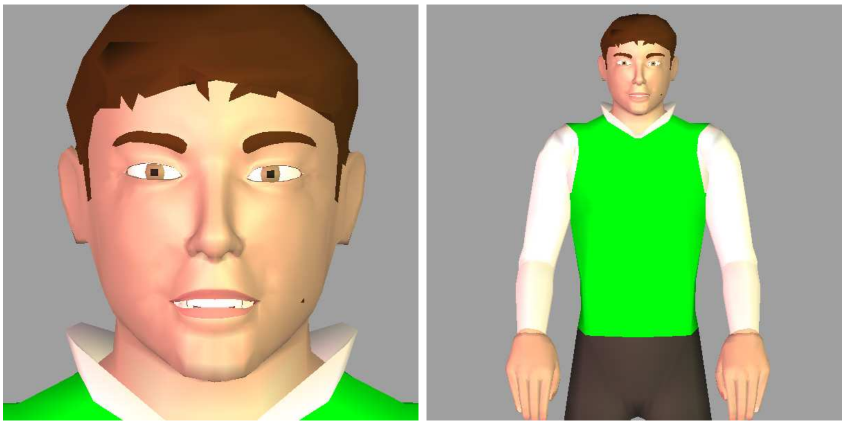
$p < 0.02$. As expected, it is due to listeners are more likely to understand better what a talking head is speaking when their distance respect to the talking head is as small as possible.

8 Conclusions and Future Work

We have presented a talking head integrated into our SpSL synthesizer. In this way, deaf people can access to the information stored by means of speech by reading the lip movements produced by the talking head.

We have also explored two different camera shots and have shown that the *Near camera* shot achieves a significant improvement compared with the *Far camera* shot in the final user evaluation. However, we have stated the *Far camera* shot is the standard visualization shot for the SpSL avatar. Further work will focus on improving the perceived quality for this *Far camera* shot. In this way, a Spherical Cubic interpolation technique will be considered.

This work has been evaluated by hearing people. Nowadays, the same tests, together with lip reading comprehension tests, are being done by hard of hearing and deaf people to test its real usability. In addition, in order to improve the usability of the talking head, new evaluation metrics will be used. We will focus on the words correctly identified by this new group of testing people, which will allow them to access to the speech-based content.



(a) Near shot (b) Far shot

Fig. 5. This figure shows the different camera shots used in the Spanish talking head evaluation.

9 Acknowledgements

Authors would like to acknowledge the FPU-UAM program for its financial support.

References

1. Nakamura, S., Yamamoto, E.: Speech-to-lip movement synthesis by maximizing audio-visual joint probability based on the em algorithm. Journal of VLSI Signal Processing Systems **27**(1-2) (February 2001) 119–126
2. Moubayed, S.A., Smet, M.D., Hamme, H.V.: Lip synchronization: from phone lattice to pca eigen-projections using neural networks. In: Proc. of Interspeech. (September 2008) 2016–2019
3. Yamamoto, E., Nakamura, S., Shikano, K.: Lip movement synthesis from speech based on hidden markov models. Speech Communication **26**(1–2) (April 1998) 105–115
4. Englebienne, G., Cootes, T., Rattray, M.: A probabilistic model for generating realistic lip movements from speech. In Platt, J., Koller, D., Singer, Y., Roweis, S., eds.: Advances in Neural Information Processing Systems 20. MIT Press, Cambridge, MA (2008) 401–408
5. Beskow, J., Granstrm, B., Nordqvist, P., Moubayed, S.A., Salvi, G., Herzke, T.: Hearing at home communication support in home environments for hearing impaired persons. In: Proc. of Interspeech. (September 2008) 2203 – 2206
6. Zoric, G., Cerekovic, A., Pandzic, I.S.: Automatic lip synchronization by speech signal analysis. In: Proc. of Interspeech. (September 2008) 2323
7. San Segundo, R., Montero, J.M., Macías-Guarasa, J., Córdoba, R., Ferreiros, J., Pardo, J.M.: Proposing a speech to gesture translation architecture for spanish deaf people. Journal of Visual Languages and Computing **19**(5) (October 2008) 523–538

8. 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.: Speech to sign language translation system for spanish. Speech Communication **50**(11-12) (November 2008) 1009–1020
9. López, F., Tejedor, J., Garrido, J., Colás, J.: Use of a hierarchical skeleton for spanish sign language 3d representation over mobile devices. In: Proc. of INTERACCION, AIPO (November 2006) 565–568
10. López, F., Tejedor, J., Bolaños, D., Colás, J.: Intérprete de lenguaje de signos en español multidispositivo. In: Proc. of IADIS-CIAWI, IADIS (October 2006) 293–296
11. Young, S., Evermann, G., Hain, T., Kershaw, D., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V., Woodland, P.: The HTK Book (for HTK Version 3.2). Microsoft Corp. and Cambridge University Engineering Department (2002)
12. Java Community Process: Jsr-184. mobile 3d graphics api for j2me. http://www.jcp.org/en/jsr/detail?id=184 (2005)
13. Quilis, A.: El comentario fonológico y fonético de textos. ARCO/LIBROS, S.A. (1998)
14. Moreno, A., Poch, D., Bonafonte, A., Lleida, E., Llisterri, J., Mariño, J.B., Nadeu, C.: Albayzin speech database: Design of the phonetic corpus. In: Proc. of Eurospeech. Volume 1. (September 1993) 653–656