Generating Gestures from Speech

R. San-Segundo, J.M. Montero, J. Macías-Guarasa, R. Córdoba, J. Ferreiros, J.M. Pardo.

Speech Technology Group. Dept. of Electronic Engineering. Universidad Politécnica de Madrid
ETSI Telecomunicación. Ciudad Universitaria s/n. 28040-Madrid. SPAIN.
{lapiz,juancho,macias,cordoba,jfl,pardo}@die.upm.es

Abstract

This article describes a first version of a system for translating speech into Spanish Sign Language. The system proposed is made up of 4 modules: speech recognizer, semantic analysis, gesture sequence generation and gesture playing. For the speech recognizer and the semantic analysis, we use modules developed by IBM and the University of Colorado respectively. The gesture sequence generation uses the semantic concepts (obtained in the semantic analysis) associating them to several Spanish Sign Language gestures. This association is carried out based on a number of generating rules.

For gesture animation, we have developed an animated character and a strategy for reducing the effort in gesture generation. This strategy consists of making the system generate automatically all agent positions necessary for the gesture animation. In this process, the system uses a few main agent positions (2-3 per second) and some interpolation strategies, both issues previously generated by the service developer.

1. Introduction

Speech and language technologies have always had an important relationship with their corresponding animated characters. These technologies provide them with new capabilities that improve the interface between them and the end users. The users can interact with animated agents using the common language. An important community of scientists worldwide is developing and evaluating virtual humans embedded in spoken language systems. These systems provide a great variety of services in very different scenarios [1][2][3][4]. In all the aforementioned systems, the synergy between language and virtual character technologies is due to the fact that virtual humans offer a more friendly computer-user interface. This synergy becomes stronger in our case where we want to develop a system to translate speech into gestures for deaf-mute people. In this case, the virtual agent appears as an essential part of the system. It has to represent the gestures obtained from the semantic analysis of the recognized words.

In Spain, during the last 20 years, there have been several proposals for normalizing Spanish Sign Language, but none of them has been very well received by the deaf-mute people. These proposals tend to constrain the sign language, limiting its flexibility. In 1991, MA. Rodríguez [5] carried out a detailed analysis of Spanish Sign Language showing the main characteristics. She showed the differences between the sign language used by deaf-mute people and the standardization proposals. This work has been one of the main studies on Spanish Sign Language and it has been the main reference in our work.

2. System Overview

In figure 1, we show the architecture proposed for translating speech into gestures for deaf-mute people. In this diagram, we have remarked on the 4 main modules, which carry out the 4 steps needed in the translation process: speech recognition, semantic analysis, gesture sequence generation and gesture playing. The position generation and the gesture animation modules permit the gesture animations needed by the gesture-playing module to be generated.

The first module (speech recognition) converts the speech utterances into text words. For this module, we have used the latest version of the IBM ViaVoice software for Spanish [6][7]. It is a voice recognition product that includes essential dictation, and command/control features. This module uses language and acoustic models adapted to Spanish pronunciation.

The semantic analysis module carries out a semantic evaluation of the text sentence (output of the speech recognizer) extracting the main concepts related to the application domain. For this module, we have used the Phoenix v3.0 parser developed at the University of Colorado (The Center for Spoken Language Research)[8][9][10][11]. This parser uses a context free grammar to extract the semantic concepts from the word sequence.

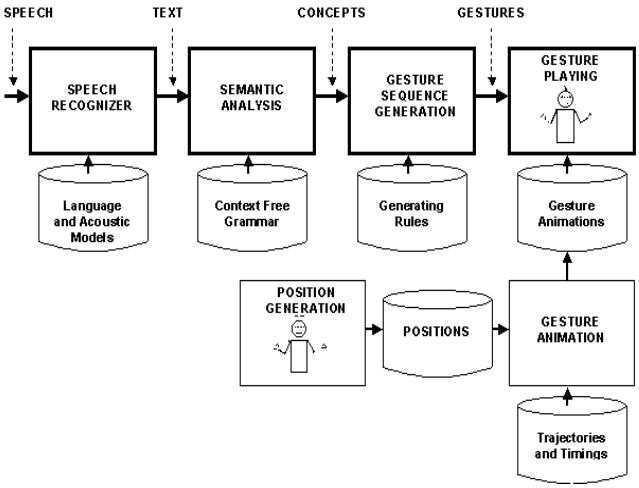

Figure 1: Speech to Gesture System Architecture.

The gesture sequence generation module processes the semantic analysis output and assigns a sequence of gestures to the semantic concepts. In this process, we consider 4 situations: one concept is mapped into a unique gesture, one concept generates several gestures, some concepts are mapped into a unique gesture, and finally, several concepts

generate several gestures. In this paper, we have studied different analyses of Spanish Sign Language [5] and we propose solutions for the 4 aforementioned situations. For solving these situations, we consider both the Context Free Grammar (semantic analysis module) and the Generating Rules (gesture generation module). The semantic analysis and the gesture generation modules are designed for restricted domain services. This means that the Context Free Grammar and the Generating Rules, used in these modules, do not contain all the possibilities for any interacting context.

In the fourth module, an animated character represents the gesture sequence. This character is a very simple representation of a human being but it permits the gestures of the sign language to be represented properly. For each gesture, the system plays a different character animation.

3. Gesture Sequence Generation

At this step, the gesture sequence generation consists of processing the semantic analysis output to obtain the final gestures that the animated agent will represent. In this process, we should point out 4 situations:

3.1. One semantic concept corresponds to a gesture

In this case, a semantic concept (parsed slot) is directly mapped onto a specific gesture. The translation is simple and it consists of assigning one gesture to each semantic concept. This gesture can be unique, independent of the word string or can be different according to the word string, which generated the concept (figure 2).

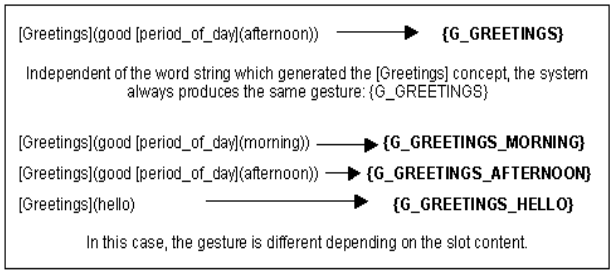

Figure 2: Assigning an unique gesture to a semantic concept.

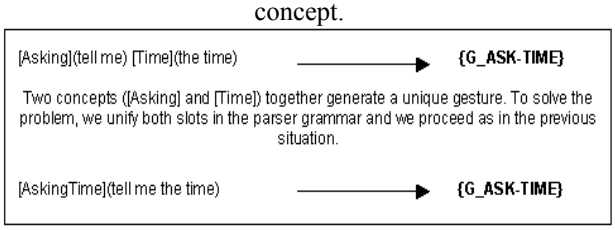

Figure 3: Assigning an unique gesture to several semantic concepts.

3.2. Several semantic concepts are mapped onto a unique gesture

The second situation appears when several concepts generate a unique gesture. This situation should be solved in the previous step (semantic analysis). The solution is to unify the slots in the parser grammar (resulting in just one slot) and to proceed as in the previous situation (figure 3).

The Phoenix parser (semantic analysis) provides the possibility of organizing the slots (concepts) into a hierarchical structure. This fact allows us to establish more complicate relationships between them in order to generate a unique gesture. As in the previous situation, the gesture to generate can be different according to the slot content or not.

3.3. One semantic concept (slot) generates several gestures

The third situation occurs when it is necessary to generate several gestures from a unique concept. This problem strongly justifies the need for this module: the gesture sequence generation module. Similar to previous sections, the gesture sequence and its order could depend on the concept and its content, or just on the concept. This situation appears in many translation issues:

- VERBS. A verb concept generates a gesture related to the action proposed by the verb and a gesture providing information about the action term (past, present or future), the action subject and gerund action (figure 4).

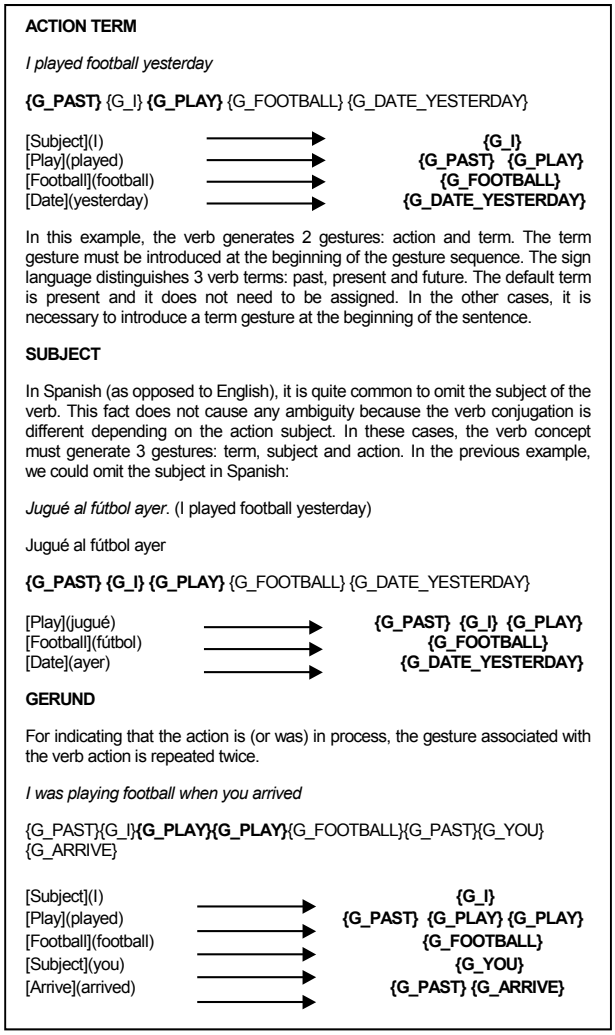

Figure 4: Type of gesture sequences generated by verb concepts.

- GENERAL and SPECIFIC NOUNS. In sign language there is a tendency to refer to things with high precision or concretion. As a result of this, there are several domains where several specific nouns exist, but there is

no general noun to generally refer to them. For example, this fact happens with the metals: there are different gestures to refer to gold, silver, copper,... but there is no general gesture to refer to the concept of metal. The same thing happens when considering furniture: there are several gestures for table, chair, bed, etc. but there is no general gesture referring the concept of furniture in general. This problem is solved in sign language by introducing several specific gestures (figure 5):

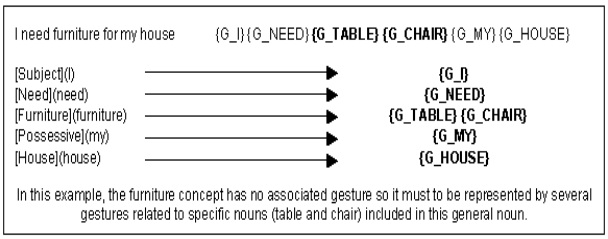

Figure 5: Gestures for general nouns not presented in the sign language.

3.4. Several semantic concepts (parsed slots) generate several gestures

Finally the most complicated situation appears when it is necessary to generate several gestures from several concepts with certain relationships between them. Some examples are the followings:

- Verb/Action gesture depends on the subject. For example, the verb "fly" is represented with different gestures depending on the action subject: bird, plane, butterfly, etc.
- A similar situation crops up when the gesture associated to an adjective changes depending on the qualified object. For example, the gesture for the adjective "good" is different when referring to a person or a material thing.

These cases are less frequent than the other ones. In our system, they are solved by mixing the strategies carried out in sections 3.2 and 3.3: first, we group the different concepts under a unique slot structure, and then we apply similar strategies as in the section 3.3, to generate a gesture sequence from a unique semantic concept structure. The characteristics of the sign language used by Spanish people have been extracted from [5] where we obtained an extended and detailed description.

4. Gesture Animation

In order to represent the gesture sequence (generated in the previous module), we have developed an animated character. This character is a simple representation of a human person but it is detailed enough to represent the gestures used in sign language. In this section, we focus on the description of this character and gesture generation. One of the main issues dealt with in this section is the way to generate gesture animations from a very small number of character positions to reduce drastically the effort in gesture generation.

4.1. The Animated Agent: AGR (Agent for Gesture Representation)

For representing the gestures, we have developed a very simple animated agent. This agent is made up by combining rectangles, circumferences and different sized lines (figure 6).

AGR is made up of 5 fixed points (head circumference and trunk rectangle) and 60 mobile points: 18 for the right arm, hand and fingers, 18 for the left arm, hand and fingers, and 24 for the face representation (eyes, mouth, eyebrows and two hairs).

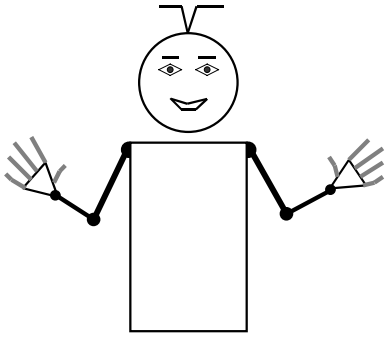

Figure 6: AGR: Agent for Gesture Representation.

In figure 7, we show the hand letter positions (dactylography).

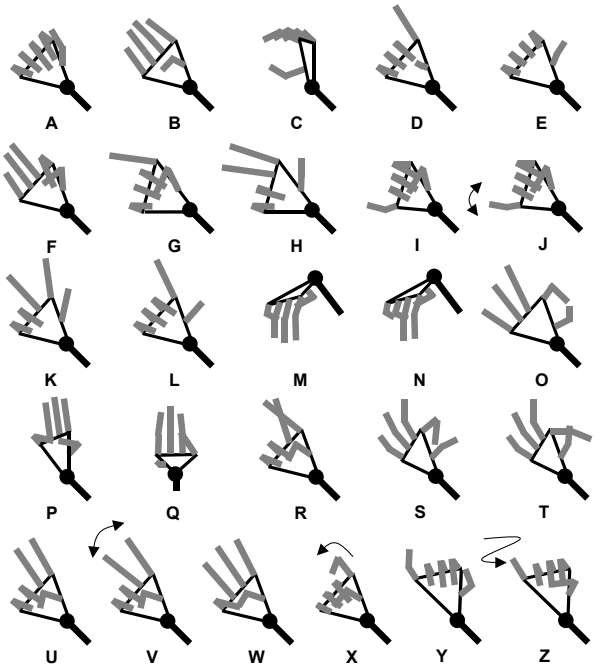

Figure 7: Signs for the letters

4.2. Obtaining Gesture Animations from Agent positions

An animation is generated automatically from a very small set of agent positions. These are defined previously using a visual tool. The main target of this module is to generate an animation using as few positions as possible in order to reduce drastically the effort of generating gesture animations. A typical gesture usually takes around 2 seconds. This means that, considering 20 frames per second (for a continuous movement), we should create 40 frames/agent position for a typical gesture. In sign language, there are more than 5000 different gestures. Creating their animations is very hard work. In this paper, we propose a strategy that reduces this effort. The main idea is to define a small number of frames/positions (around 4-5 per gesture), and to generate the

intermediate positions automatically. The program creates these frames by interpolation. For any subsequence (positions created automatically between two positions defined by the user), the user can specify different interpolations. In order to design an interpolation, it is necessary to define two aspects: the trajectory and timing.

- The user can define the trajectory that any point of the agent body will follow when moving from the initial to the final position (figure 8). The trajectory is specified by the user in a visual interface (with an adequate zoom) by moving the mouse cursor. When the user defines a trajectory, this trajectory can be assigned to a unique mobile point, a set of mobile points, or to all mobile points. For the mobile points for which the user does not define any trajectory, the program generates a rectilinear one by default. This fact allows a complete specification. The default trajectory is not fixed and can also be modified by the user.
- The second aspect to define is the timing: how fast the point passes through the different parts of the trajectory. The trajectory is a continuous line (infinite points) but the number of intermediate positions is small: around 10. Because of this, the user needs to specify where, in the trajectory, the mobile point will be situated for each of the interpolated positions. In the same visual interface (figure 8), several intermediate circles appear, as many as intermediate positions. The user can position each circle at any trajectory point (as it is not possible to change the circle order). When the user defines timing, it is associated to a unique mobile point, a set of mobile points or to all mobile points. By default, if no timing is specified, the program positions the intermediate points equidistantly.

Two points with the same trajectory can have different timings. The interpolated positions/frames are created by the program combining the trajectory and timing associated to each point. In this process, the program checks some limitations concerning the length of some parts of the body. The goal is to avoid generating extremely deformed gestures. It also checks some conditions: eyes, eyebrow and mouth must be within the head limits, or the pupil should be situated within the eye limits.

5. Conclusions

In this paper, we have proposed an architecture for a speech-to-gesture translator made up of 4 modules: speech recognizer, semantic analysis, gesture sequence generation and gesture sequence animation (gesture playing). The main effort in this work has been focused on the gesture sequence generation and gesture animation. The gesture sequence generation is applied over the semantic analysis provided by the Phoenix parser. The hardest situation has been when a semantic concept generates several gestures. For this case, the detailed description of the Spanish Sign Language, carried out by M.A. Rodríguez [5], has been very useful.

For the gesture animations, we have developed an animated agent and a strategy for reducing the gesture generation time. This strategy consists of combining agent positions created by the user and positions generated automatically by the system. The position generation is carried out by interpolation considering point trajectories and timings designed by the user.

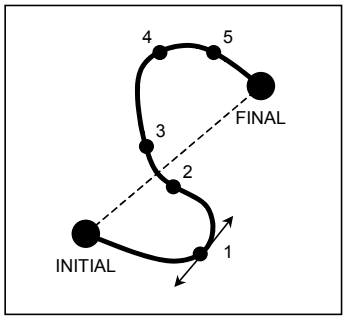

Figure 8: Timing specification with 5 intermediate positions

6. References

[1] Cassell, J., Stocky, T., Bickmore, T., Gao, Y., Nakano, Y., Ryokai, K., Tversky, D., Vaucelle, C., Vilhjálmsson, 2002. "MACK: Media lab Autonomous Conversational Kiosk" in Proc. Of Imagina: Intelligent Autonomous Agents, Monte Carlo, Monaco, 2002.

[2] Gustafson, J., 2002. "Developing multimodal spoken dialogue systems- Empirical studies of spoken human-computer interactions". Ph D. Dissertation. Dept. Speech, Music and Hearing, Royal Inst. of Technology, Stockholm, Sweden, 2002.

[3] Granström, B., House, D., Beskow, J., 2002. "Speech and Gestures for Talking Faces in Conversational Dialogue Systems" Multimodality in Language and Speech Systems. Kluwer Academic Publishers. pp 209-241. 2002.

[4] Cole, R., Van Vuuren, S., Pellom, B., Hacioglu, K., Ma, J., Movellan, J., Schwartz, S., Wade-Stein, D., Ward, W., Yan, J., 2003. "Perceptive Animated Interfaces: First Steps Toward a New Paradigm for Human Computer Interaction", in IEEE Transactions on Multimedia: Special Issue on Human Computer Interaction, vol. 91, no. 9, pp. 1391-1405. 2003.

[5] Rodríguez, M.A. 1991. "Lenguaje de signos" Phd Dissertation. Confederación Nacional de Sordos Españoles (CNSE) and Fundación ONCE. Madrid. Spain. 1991.

[6] IBM web: http://www.ibm.com/

[7] Outsource-sl web: http://www.outsource-sl.com/fabricantes/IBM/ViaVoiceStd.htm

[8] Ward, W., 1994. "Extracting Information From Spontaneous Speech" International Conference on Spoken Language Processing. September. 1994.

[9] Ward, W., Pellom B. 1999. "The CU Communicator System." Proc. IEEE Workshop on Automatic speech Recognition and Understanding (ASRU), Keystone Colorado. 1999.

[10] Ward, W., Pellom, B. 2002. "The Phoenix Parser User Manual" downloadable from http://communicator.colorado.edu.

[11] Phoenix Parser Software. http://communicator.colorado.edu.