Phonemicization for the Generation of Vi-syllable

José Antonio Radilla-Ávila, María Lucila Morales-Rodríguez, Arturo Hernández-Ramírez, J. Apolinar Ramírez-Saldivar, Juan J. González-Barbosa.

Instituto Tecnológico de Cd. Madero, México
E-mail: radilla_21@hotmail.com, lmoralesrdz@gmail.com, ahr@prodigy.net.mx apolinar_r@yahoo.com, jjgonzalezbarbosa@hotmail.com

Abstract. This work presents the application of linguistic rules for the creation of vi-syllables of phrases written in the Spanish language in order to increase the believability of the labial synchronization in virtual characters. Especially, we describe the implementation of the phonological transcription and Syllabication as part of Phonemicization's process.

Keywords: Virtual characters, Visemes, Vi-syllable, Labial Synchronization, Phonological Transcription Syllabication.

1 Introduction

Our work is interested in increasing the believability of the labial synchronization in virtual characters, we present a proposal for the creation of vi-syllables of phrases written in the Spanish language applying linguistic basis to select and associate a reduced number of visemes to a syllable (vi-syllable).

The creation of vi-syllables allows us to reduce the labial movements and thus increase the believability of the lips synchronization, choosing only the most representative visemes of a syllable. Usually a syllable is composed of two or more phonemes and a viseme is the visual reproduction of the phonemes.

The aim of our proposal is to create more believable virtual character by the selection of visemes related to the prosody and phonetics of Spanish language utterances.

This paper is structured as follows: in the section two we describe the concepts related to the labial synchronization and the linguistics basis, in the section three we present the State of the Art, in the section four the proposal of our architecture is described and finally in the section five we present the conclusions and future works.

2 Labial synchronization in Virtual Characters

In this section we present the theoretical background of three main issues involved in our topic of investigation: the animation of virtual characters, the labial synchronization and linguistics.

2.1 Labial animation in Virtual Characters

A virtual character is an animated actor represented by a computer who coexists inside a virtual universe and that might interact with a user [1].

The animation is a process used to give the sensation of movement to images or drawings. In our work we seek to create animations lips to accompany the articulation of the sentences issued in Spanish, in order to increase the sense of immersion provided by the virtual character.

The lip sync is defined as the movements of the lips to choose the matching Visemes artificial human voice [2]. According to Keith Lango [3], the trend for the generation of labial synchronization is: 1) make visemes for the "sounds" (phonemes), like M, E, S, etc., 2) listen to the track of audio or the recording voice, 3) For every sound to that they listen, there is given to the "sound" a representation of a nearby or equal image to 100 %, 4) make a preview render of the lip sync, 5) watch the mouth flap out of control and 6) wonder what went wrong.

Received Oct 10, 2011 / Accepted Dec 31, 2011
The visemes are the movements of lips when is produced a certain phoneme [4], in other words a viseme is the visual representation of a phoneme (section sees 2.2).

The problem of following this trend is to be too literal in the animation of the speech, trying to animate the letters that compose the words, instead of emphasizing the major sounds needed to communicate the idea of speech. Keith Lango proposes two basic principles for the solution of this problem [3].

Principle #1: The letters are not sounds. The sounds are not letters. There are no letters in the animation of labial synchronization. Generally the beginners will create an exact reply of the face saying a ‘letter’ in isolated form. An example of this, it is when the sound of a letter is said alone, this one is not like when it is accompanied of another letter (consonant or vowel), for example the "t" alone it sounds as "ti", the form of mouth is different when the "t" is accompanied of a consonant for example in "trompo".

Principle #2: The mouth shapes for sounds must be animated in context. The sound shape that precedes it influences the shape of current sound. Of equal way, the following shape of sound is anticipated by the shape of current sound. When somebody says a letter, for example the "U", the mouth is already in the shape of "U", before this one is listened.

2.2. Linguistics for the labial synchronization

In our work we use two disciplines of the linguistics: the phonetics and phonology. The phonetics, studies the sounds of the language, that is to say the allophones. The phonology is the discipline that represents these sounds from the point of view of the language, that is to say, the phonemes [8].

A phoneme (writes between / /) is the abstract and widespread representation of a sound with the capability of differentiate meanings [8]. At the moment of classifying the phonemes, two principal criteria are used: The place of articulation, which is the place in the mouth where a particular phoneme is pronounced and the manner of articulation, that is to say, the processes that are started to enounce a phoneme [9].

Depending on the place of articulation, the phonemes could be classified into:

Bilabial: participation of lips, the superior and the low one.
Labiodental: low lip with top teeth.
Linguointerdental: the tongue between the teeth.
Linguodental: the tongue touches the back part of the top teeth.
Linguopalatal: the tongue touches the top alveoli.
Linguovelar: the tongue touches the veil.
Uvular: the posterior part of the tongue is placed against or near the uvula, that is to say, more behind in the mouth that the veil.

The vowels differ to each other by their place of articulation which is characterized by two attributes (sees figure 1) [10]: the opening or distance that exists between the palate and the tongue, and the location or place of the mouth in which we place the tongue when we speak [10].

In the pronunciation of a language, it is necessary to consider the physiological elements that will allow us to describe the phonemes, that is to say, the nature of the obstacle on the exit of the air and the processes involved in the phoneme articulation. The physiological elements involved are [9]. The vowels chords, which it provokes that a consonant is sonorous (it vibrates) or deaf (it does not vibrate). The tongue, which rubs parts of the mouth cavity. The veil of the palate, the palate, the alveoli, the teeth, the lips the nasal cavity (only the nasal consonants).
According to their manner of articulation the consonantal phonemes can be classified into 6 types:

**Occlusive:** They are consonants characterized by a total blockade of the flow of the air caused by a complete obstruction that is created when an active articulator does total contact with a passive articulator. An explosion is produced.

**Fricative:** They are consonants that are articulated forcing the air across a narrow crack created by the friction between two articulators but without the flow of the air is interrupted. A fret is produced.

**Affricates:** They are consonants which joint includes a phase of total obstruction followed by a phase of friction. During the phase of total obstruction the flow of the air is interrupted momentarily whereas during the phase of friction the air escapes forcibly.

**Nasals:** They are consonants which joint needs a total obstruction in the oral cavity accompanied of a decrease, allowing that the air should flow across the nasal cavity.

**Sidewise:** They are consonants in which the tongue produces a central blockade but the air escapes laterally.

**Vibrant:** They are consonants characterized by a vibratory movement of the active articulator produced for the vibration of the tongue.

The figure 2 present the classification of the consonantal phonemes of the Spanish based on its manner and place of articulation.

Each one of the variants that are given in the pronunciation of the same phoneme is called allophones (represented by brackets []). The variants are related to the position of phoneme in the word or syllable, or according to the character of the neighboring phonemes [5]. The figure 3 shows the classification of the consonantal allophones in Spanish.
2.3 Phonological Transcription

When we want to move the phonemes to the writing, we use the graphical representations to which we call normally letters or, graphemes [5, 6]. A grapheme is the minimal distinctive unit of a system of writing, or, the minimal element by which two words can be distinguished in writing in a language. It is represented between <> [7].

The rules of transcription of grapheme to phoneme, converts a grapheme to a phoneme [5]. This type of transcription represents only the phonic elements (phonemes) provided with linguistic function. In our work we use 27 rules of conversion grapheme to phoneme proposed by Ríos-Mestre [12], in the figure 4 appear the first 5 rules.

<table>
<thead>
<tr>
<th>Rule</th>
<th>Grapheme</th>
<th>Procedure</th>
<th>Phoneme</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>&lt;c&gt;</td>
<td>, , , , c., 'ééíí, , , , ,</td>
<td>/k/</td>
<td>casa, copa, cuadro, clase</td>
</tr>
<tr>
<td></td>
<td></td>
<td>k0.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>&lt;c&gt;</td>
<td>, , , , c., 'ééíí, , , , ,</td>
<td>/z/</td>
<td>cena, cine</td>
</tr>
<tr>
<td></td>
<td></td>
<td>z0.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>&lt;g&gt;</td>
<td>, , , , g., 'ééíí, , , , ,</td>
<td>/x/</td>
<td>gente, girasol, general</td>
</tr>
<tr>
<td></td>
<td></td>
<td>x0.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>&lt;h&gt;</td>
<td>...., h., , , , ,</td>
<td>/Ø/</td>
<td>Hueso, hueco huevo</td>
</tr>
<tr>
<td>5</td>
<td>&lt;i&gt;</td>
<td>, , , , , , , , , aáééíóóúú'-, h., l., aeou, , , , ,</td>
<td>/y/</td>
<td>Mohiento, rompehielos</td>
</tr>
</tbody>
</table>

Figure 3. Classification of the consonantal allophones of the Spanish according to its manner and point of joint.

Figure 4. Fragment of the rules of phonological Transcription Rios-Mestre's (grapheme-phoneme) [12].
2.4 Syllabification and Phonetic Transcription


The process of syllabification is composed of three stages: the pre-syllabification, syllabification and the syllabic adjustment, which use a mark of syllabic division script (-), with the purpose of being able to represent a phrase in a syllable.

The rules of pre-syllabification insert a mark of syllabic division in those words that will be considered to be exceptions of the rules of syllabication for its posterior use in the phonological transcription and in the process of syllabication. We present a fragment of the rules of pre-syllabification in Figure 5.

FON0.REG “Rules of Pre-syllabification”

<table>
<thead>
<tr>
<th>Rule</th>
<th>Entry</th>
<th>Procedure</th>
<th>Exit</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>VOL</td>
<td>s., aou., b., lir., ...,</td>
<td>VO-L</td>
<td>sub-lingual abrogativo ob-repticio</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.=0.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>CCV</td>
<td>_ds., _ei., sn., h., ...,</td>
<td>C-CV</td>
<td>des-hielo des-hierbo des-hierro des-hueso des-hielo des-hielas</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.=0.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>VCV</td>
<td>.., aeiodu., h., ...,</td>
<td>V-CV</td>
<td>cuenta-hilos barbihecho cariharto</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.=0.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 5. Fragment of the rules of pre-syllabification according to Ríos-Mestre [12].

The rules of syllabification have the purpose of separating with a syllabic mark (-) two or more syllables. There exist 18 rules of syllabication that classify two big groups, the consonants and the vowels. The first twelve rules of the syllabification correspond to the consonants and the six remaining rules correspond to the vowels.

FON2.REG “Rules of Syllabification”

<table>
<thead>
<tr>
<th>Rule</th>
<th>Entry</th>
<th>Procedure</th>
<th>Exit</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>VCV</td>
<td>...aáeéiíoóúú., ptkbdygfzsmnñílrRw., aeiou., ..,</td>
<td>V-CV</td>
<td>cine, peine, cosas, amo</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.=0.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>VCV</td>
<td>.., aeiou., ptkbdygfzsmnñílrRw., aéióú., ..,</td>
<td>V-CV</td>
<td>casa, aze, piso, ocupar, sucio</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.=0.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>{ y _ }</td>
<td>.., aééíioóúú., ptkbdygfzsmnñílrRw., y., ..,</td>
<td>{ - y _ }</td>
<td>Curry, ferry, rally, Vichy</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.=0.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 6. Fragment of the rules of syllabification according to Ríos-Mestre [12].

The Syllabic adjustment carries out a syllabic division between two syllables in order to correct the foreign words, for example /án-g-s-trom/# /ans-trom/ with the purpose of eliminating the phoneme that does not have a phonetic representation in our language.

The module of syllabic adjustment consists of seventeen rules that are divided in two groups where the first eleven rules correspond to the elimination of the phonemes that does not have a phonetic interrepresentation in the Spanish language. The six remaining rules correspond to the elimination of the syllabic sign named script (-). In Figure 7 we present a fragment of the rules corresponding to the module of syllabic adjustment.
According to Ríos-Mestre, after the division of syllables (syllabification) the process continues with phonetic transcription, that apply rules of transcription phoneme - allophone to the syllables detected in relation to the pronunciation. The result is the graphical representation of the sounds of the speech, using a phonetic alphabet [5, 8].

### 3 State of the Art of the Labial Synchronization

In this section they present four works related to the labial synchronization, we focus them in two principal issues, the selection of visemes and the creation vi-syllables.

#### 3.1 Works related to Visemes's Selection

Ferreira and De la Rosa [13] in their work “An offer of integration of facial animation and synthetic voice” presents the analysis, evaluation and study of feasibility of the integration of the standard MPEG-4 with the utilization of tools of Text to Speech, which allows the translation of text written to synthetic voice. The process of selection of visemes uses the standard MPEG-4 that contains a complex set of parameters FDP (Facial Definition Parameters), used for the standardization of the face. The FAPs are expressed in FAPUs's terms (Facial Animation Parameter Units) that consist of the representation of the distance between the different typical points defined to preserve a proportionality adapted to the size of the geometric model.

A fundamental characteristic of work of Ferreira and De la Rosa is the synchronization of the animation with the synthetic voice generated by the engines of translation from a written text, using asynchronous functions of reproduction of means provided by the APIs of programming of the most used operating systems.

In 2009, Ceballos et al. [14] presented the work, “Seguimiento del contorno externo de la boca en imágenes de video”. The aim of this investigation is to propose an algorithm for the follow-up of the external contour of the mouth; this work is based on the appearance and on morphologic restrictions defined in the standard MPEG-4, using pixels for the follow-up of the external contour of the mouth in images to color. 68 FAPs are defined in the standard MPEG-4, which they divide in 10 groups corresponding to the recognition of the speech, generally the group’s use 2 and 8, which describe the movement of the internal and external contour of the mouth respectively.
The authors present an algorithm that uses the group 8 that the external contour of the lips describes. This algorithm divides in four steps for a better follow-up of the external contour of the mouth which are: Detection of the region of the mouth, location of the lips (initialization), follow-up of the lips and the extraction of characteristics with the purpose of locating ten points of the external contour of the mouth for the first picture of video. Once found the points in the sequence of video, for the applications of the follow-up of the lips, the characteristics of the form of the mouth must be calculated. Having 10 points on the whole sequence of video, it is possible to find, between many an others the area of the region inside the lips, the roundness, the shape factor, the relation between the horizontal axis and the vertical one, the perimeter and different geometric relations between the points. Their principal bounding is that they not achieved to establish a methodology adapted to realize precise follow-up of the form of the real time lips.

In 2007 Melenchón [15] presented his doctoral thesis called “Síntesis Audiovisual Realista Personalizable”. This work uses genetic parallel algorithms to choose the set of visemes and their facial animation. The information of the allophones and visemes was made for the Spanish language of Spain.

### 3.2 Work related to the creation of vi-syllables

Kshirsagar [16] presents the work “Vi-syllable Based Speech Animation”, where implements vi-syllables (the representation of a syllable like a viseme) for the animation of an English TTS voice. In this work, the first phase is the arbitrary obtaining of syllables of a phoneme stream, later, these syllables are checked by phonological rules for his approval, the final syllable stream is later used with a database of visyllables recorded using a facial motion capture system (see Figure 9).

![Figure 9. Animation of voice based in real time vi-syllables [16].](image)

### 4 Proposal of architecture

In this work we propose to automate the syllabification process and the creation of vi-syllables of phrases written in the Spanish language in order to increase the credibility of the labial synchronization in virtual character. Our general aim is the creation of vi-syllables of phrases written in the Spanish language (from México) applying linguistic basis to accomplish the process of syllabification. To achieve this objective, we carried out a bibliography research related to methodologies for phonetic and phonological transcription of Spanish words, as well as the automation process of syllabification in Spanish.

Our proposal is based on Rios-Mestre's work [12], which presents five modules for the process of phonemization, which are: 1) Rules of transcription grapheme-phoneme (FON1.REG), 2) Rules of syllabification (FON2.REG), 3) Rules of syllabic adjustment (FON3.REG), 4) Rules of accentuation (FON4.REG) and 5) Rules of phonemic adjustment (FON5.REG).

Our architecture uses the first three modules of the Rios-Mestre's Phonemization Process, and consists of four blocks (see figure 10): 1) Application of Phonemization's Rules, 2) Transcription to place of articulation, 3) selector of vi-syllables and 4) Animation of vi-syllables (visemes associated with the syllables).
Figure 10. General scheme of the process of synchronization of the visemes

For the application of Phonemization's rules the entry is an utterance, the result of the application of the rules is the chain pre-syllabifying. This string is going to do the entry of the process of phonological transcription and as result it will give us the string of phonemes, this string of phonemes will be in use in the process of transcription to the place of articulation and as entry to the rules of syllabification, which will give a syllabifying string (see Figure 11).

The next process, the Transcription by place of articulation represents the phonemes selected according to its place of articulation (see section 2.2). The result of this transcription will be the entry for the process of generation of vi-syllables.

The last phase of our process is the Generation of vi-syllables, which is composed by the module of selection and animation of visemes. This phase will allow us to select the visual representation of the syllable that match and represent the animation of visemes. The selection process based on data mining uses the string of phonemes syllabifying, the place of articulation transcription and a Corpus of visemes.

To illustrate the operation of the architecture (see Figure 11), we present an example with the word `deshielo`, we describe the steps applied of each one of the modules presented in Phonemicization Rios-Mestre's process.

Step 1: To obtain the phrase to processing: `deshielo`.
Step 2: To apply rules of pre-syllabification: the rule applying for the example is the rule 2 (see Figure 12)
With the application of these two steps, the chain pre-syllabifying of the word deshielo contain a mark of syllabic division between the graphemes s and h with the purpose of use that in the conversion of graphemes to phonemes.

<table>
<thead>
<tr>
<th>Example</th>
<th>Entry</th>
<th>Procedure</th>
<th>Exit</th>
<th>Rule 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>deshielo</td>
<td>CCV</td>
<td>_ds,, _ei,, sn,, h,, ,, ,, ,, -0,=0.</td>
<td>C-CV</td>
<td>des-hielo</td>
</tr>
</tbody>
</table>

Step 3: Once pre-syllabifying the word deshielo, continues the application of the phonological transcription rules to each one of the grapheme of the word. According to the rules of transcription, the graphemes d,e,s,e,l,o are directly phonemes for their similarity between the grapheme and phoneme. However, the rule 5 is applied: the grapheme h accompanied on its right on the grapheme e is converted to phoneme y (see Figure 13).

<table>
<thead>
<tr>
<th>Example</th>
<th>Entry</th>
<th>Procedure</th>
<th>Exit</th>
<th>Rule 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>des-hielo</td>
<td>i</td>
<td>,, _aæeioouu-,, h,, i,, aeou,, ,, ,, y0.</td>
<td>y</td>
<td>des-yelo</td>
</tr>
</tbody>
</table>

Step 4: We continues applying the rules of syllabification, that is to say, the separation in syllables of two or more phonemes. There will insert a mark of syllabic division with the application of the rule 1 of the Syllabification between the phonemes e and l. It is possible to notice that already there exists a mark of syllabic division between the phonemes s and y for the application of the rules in Pre-syllabification's process (see Figure 14).
Step 5: The next step is the application of the rules of syllabic adjustment with the purpose of correcting some mistakes during the syllabification. The string des-ye-lo does not need any syllabic adjustment.

Step 6: The last stage applies the rules of transcription to place of articulation, that is, the classification of the place in the mouth (shape of lip, place tongue, etc.) when we produce a certain phoneme. The results of the process are presented in the Figure 15.

<table>
<thead>
<tr>
<th>Example</th>
<th>Phoneme</th>
<th>Classification</th>
</tr>
</thead>
<tbody>
<tr>
<td>des-ye-lo</td>
<td>d</td>
<td>Dental</td>
</tr>
<tr>
<td></td>
<td>e</td>
<td>Front – Mid</td>
</tr>
<tr>
<td></td>
<td>s</td>
<td>Alveolar</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>y</td>
<td>Fricative</td>
</tr>
<tr>
<td></td>
<td>e</td>
<td>Front – Mid</td>
</tr>
<tr>
<td></td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>l</td>
<td>Alveolar</td>
</tr>
<tr>
<td></td>
<td>o</td>
<td>Back – Mid</td>
</tr>
</tbody>
</table>

Figure 15. Example of the transcription to place of articulation.

5 Conclusions and Future Works

This work presented the concepts related to the labial synchronization in virtual characters and the linguistics background that we used for the application of linguistic rules for the creation of vi-syllables of phrases written in the Spanish language in order to increase the believability of the labial synchronization in virtual characters.

We describe the general scheme of the process of Phonemicization of Ríos-Mestre and the integration of the phonological transcription and Syllabification processes as part of our architecture. The result of these processes will be used into a data mining process in order to select the viseme adapted to the phonemes of a syllable, finally the process concludes with the animation of the vi-syllables selected.

This work is in development and we are working in determine the data mining technique that will be used to solve the selection of the visemes that will allow the creation of vi-syllables.

6 References