Audio-driven Emotional Speech Animation

C. Charalambous, Z. Yumak and A. F. van der Stappen
Utrecht University, Department of Information and Computing Sciences, the Netherlands

Figure 1: Phoneme-to-viseme mapping: “hello” -> h @ l @ U -> GK AHH L OHH UUU

1. Introduction

Nowadays, games require a high level of realism to create a bond between the player and the game characters. The bond is strengthened with the quality of the facial animation. Different approaches to lip-synchronized speech animation were developed over the years. Procedural approaches are a better choice in terms of the control of the animation while they might not reach the level of naturalness in performance-capture [WBLP11] [CWW16] and data-driven approaches [KAL17] [TKY17]. Our goal is to develop an audio-driven speech animation method for interactive game characters where the control aspect is high priority. While doing this, we want to push the boundaries of naturalness by introducing the effect of emotions. Although there are various approaches to procedural speech animation, they do not explicitly take into account the effect of emotions on the mouth movement. Recently, Edwards et al. [ELFS16] introduced the JALI model to simulate different speech styles, although they did not analyze specific emotion categories such as happy, sad and angry.

Our contribution is two-fold: 1) An expressive speech animation model that takes into account emotional variations in audio; 2) A co-articulation model based on dynamic linguistic rules varying among different emotions. Figure 2 shows the overall pipeline.

An audio file with its corresponding transcript is given as input to a phoneme extraction tool and the extracted phonemes are mapped to their visual counterparts, the visemes. The audio is analyzed, pitch and intensity values are mapped to the weight of the visemes. Co-articulation rules are applied to generate the final animation. In the following sections, we present the expressive speech and the co-articulation model and an experiment that shows the results. Finally, we point out the limitations and future work.

2. Expressive speech

Pitch and intensity influence all vowels as well as several consonants. High mean fundamental frequency, which is the metric for pitch, and high mean decibel values, which is the metric for intensity, are correlated with high arousal [SAHS05]. High arousal occurs during intense emotions such as joy or anger, where stronger articulation is observed. On the other hand, lower arousal is observed with emotions such as sadness or boredom. To calculate the weight of the visemes, we define Equation (1) and (2). For vowels we take the average of weight $W_F$ and $W_I$, since vowel articulation is influenced by both frequency and intensity. For the consonants (plosives and fricatives), we only take into account the pitch influenced weight $W_F$.

$$W_F = \begin{cases} (F_{\text{max}} - F) \times \frac{W_{\text{max}}}{F_{\text{max}}} + W_p, & \text{if } F \geq \overline{F} \\ (\overline{F} - F_{\text{min}}) \times \frac{W_{\text{max}}}{\overline{F} - F_{\text{min}}} + W_{\text{min}}, & \text{otherwise} \end{cases}$$ (1)

$$W_I = \begin{cases} (I_{\text{max}} - I) \times \frac{W_{\text{max}}}{I_{\text{max}}} + W_p, & \text{if } I \geq \overline{I} \\ (\overline{I} - I_{\text{min}}) \times \frac{W_{\text{max}}}{\overline{I} - I_{\text{min}}} + W_{\text{min}}, & \text{otherwise} \end{cases}$$ (2)

Denoted with $F_p$ and $I_p$ are the mean individual frequency and intensity values for the respective phoneme $p$, while $W_p$ is the mean weight. $F_{\text{min}}, I_{\text{min}}$ and $F_{\text{max}}, I_{\text{max}}$ are the minimum and maximum frequency and intensity values for the whole audio clip while...
$W_{\text{min}}$ and $W_{\text{max}}$ represent the minimum and maximum weight of the viseme. Male frequency varies from 85Hz to 180Hz, while female frequency varies from 165Hz to 255Hz [BO00]. During our experiments we noticed that these values were not always observed in the audio clips and therefore they needed manual adjustment.

2.1. Co-articulation

Another important aspect of speech is co-articulation. A viseme starts and ends before its corresponding phoneme. $W_{\text{onset}}(t_1)$ indicates the viseme's weight for phoneme $P_i$ at time $t_1$. The weight of the viseme continues increasing until it reaches the apex of its phoneme and then starts decreasing until it reaches zero [IMG*04]. $D_P$ and $D_V$ denotes for the duration of the phoneme and the viseme (Figure 3). It is known that speech rate can strongly influence the way people speak and therefore articulate their lip muscles. With fast speech rates, the duration of and the distance between the phonemes is minimal. Hence, defining static onset/offset values cause the visemes overlap each other extensively, creating misleading and unnatural speech animation. To counter this problem, we make both the onset and the offset intervals varying. We added an influence parameter $c$ which is multiplied with the tonset and the toffset. The influence parameter $c$ is applied in different conditions; vowel over vowel, vowel over consonant, consonant over vowel and consonant over consonant [Wal82]. We experimented with different values and selected an optimal value for each condition. Based on the findings from Taylor et al. [TT14], we also added phoneme substitution or deletion rules. We define a phoneme duration threshold, in which several consonants are dropped from speech (i.e. $h$, $t$). In addition, phonemes that contain more than one letter, known as di- phoneme and then starts decreasing until it reaches zero [IMG*04].

$$t_{\text{offset}}(P_i) = D_{P_i} * 0.75$$

**Figure 3: Co-articulation scheme**

3. Experiment, Results and Future Work

We conducted an experiment to evaluate our method with two commercial speech animation tools: Face-FX and RogoDigital. We generated 34 videos representing various emotional states (happy, sad, angry, song, poem) with varying male/female voices (3 voices per gender). The audio files were obtained from Ravdess database and LibriVox. The users watched two videos each time and selected the best one. We had 26 participants, of which 16 were male and 10 were female.

<table>
<thead>
<tr>
<th>Emotion</th>
<th>Comparison 1</th>
<th>Comparison 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Happy</td>
<td>Ours 135</td>
<td>FaceFX 73</td>
</tr>
<tr>
<td>Sad</td>
<td>116</td>
<td>92</td>
</tr>
<tr>
<td>Angry</td>
<td>111</td>
<td>97</td>
</tr>
<tr>
<td>Song</td>
<td>125</td>
<td>83</td>
</tr>
<tr>
<td>Poem</td>
<td>23</td>
<td>29</td>
</tr>
</tbody>
</table>

| Female   | 259          | 183          | 231        | 211        |
| Male     | 251          | 191          | 273        | 169        |

**Table 1: Results**

The results show that our model overall scores better in happy, sad, angry and song categories, while it scored lower in the poem category (Table 1). We speculate that it is due to the long duration of the clip which has more emotional variations. Onset/offset and influence parameters for co-articulation were tested empirically and that requires further analysis. Moreover, we plan to consider more facial parameters in addition to lip motion: cheeks, jaw and nose are the facial features that are heavily influenced by speech. We also plan to compare our method with the procedural model JALI [ELFS16] and recent audio-driven methods that use deep learning [TKY*17]. Finally, we plan to extend our work to other languages.

References

[BO00] Baken R. J., Orlikoff R. F.: Clinical measurement of speech and voice. Cengage Learning, 2000


