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

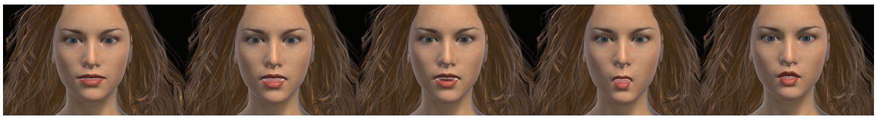

Figure 1: *Phoneme-to-viseme mapping: "hello" -> h @ l @U -> GK AHH L OHH UUU*

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] [CWW*16] and data-driven approaches [KAL*17] [TKY*17]. 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.

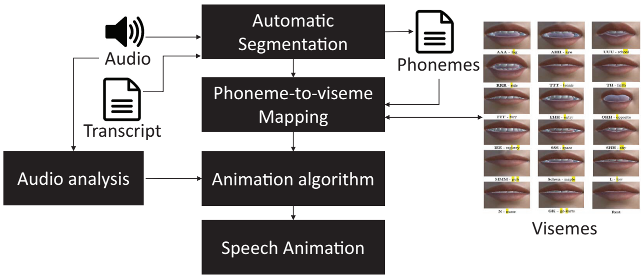

Figure 2: *Overall pipeline*

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.

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

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

Denoted with $\overline{F}_p$ and $\overline{I}_p$ are the mean individual frequency and intensity values for the respective phoneme p, while $\overline{W}_p$ is the mean weight. F_{min}, I_{min} and F_{max}, I_{max} are the minimum and maximum frequency and intensity values for the whole audio clip while

DOI: 10.2312/egp.20181019

W_{min} and W_{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. $W_{P_i}(t_c)$ indicates the viseme's weight for phoneme P_i at time t_c. 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_i} and D_{V_i} 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 t_{onset} and the t_{offset}. 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 diphones were treated differently. They were split into to two visemes which led to more variations in the mouth movements.

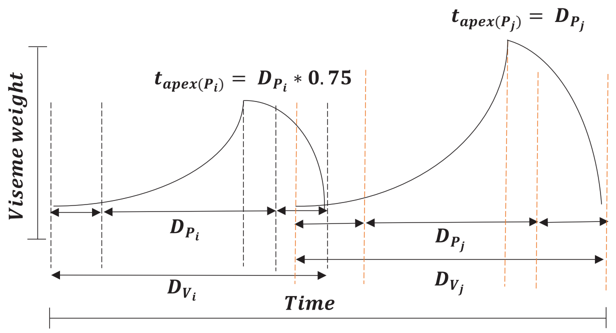

Figure 3: *Co-articulation scheme*

3. Experiment, Results and 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 1: *Results*

Emotion	**Comparison 1**		**Comparison 2**	
	Ours	FaceFX	Ours	Rogo
Happy	135	73	119	89
Sad	116	92	117	91
Angry	111	97	116	92
Song	125	83	132	76
Poem	23	29	20	32
Female	259	183	231	211
Male	251	191	273	169
Total	**510**	**374**	**504**	**380**
Total%	**58%**	**42%**	**57%**	**43%**

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. 2

[CWW*16] Cao C., Wu H., Weng Y., Shao T., Zhou K.: Real-time facial animation with image-based dynamic avatars. *ACM Trans. Graph. 35*, 4 (July 2016), 126:1–126:12. 1

[ELFS16] Edwards P., Landreth C., Fiume E., Singh K.: Jali: An animator-centric viseme model for expressive lip synchronization. *ACM Trans. Graph. 35*, 4 (July 2016), 127:1–127:11. 1, 2

[IMG*04] Ito T., Murano E. Z., Gomi H., Dimitriou M., Wolpert D. M., Franklin D. W., Ito T., Murano E. Z., Gomi H.: Fast force-generation dynamics of human articulatory muscles. *Journal Applied Physiology 96* (2004), 2318–2324. 2

[KAL*17] Karras T., Aila T., Laine S., Herva A., Lehtinen J.: Audio-driven facial animation by joint end-to-end learning of pose and emotion. *ACM Trans. Graph. 36*, 4 (July 2017), 94:1–94:12. 1

[SAHS05] Schroder M., Albrecht I., Haber J., Seidel H. P.: Mixed feelings: Expression of non-basic emotions in a muscle-based talking head. *Virtual Reality 8*, 4 (2005), 201–212. 1

[TKY*17] Taylor S., Kim T., Yue Y., Mahler M., Krahe J., Rodriguez A. G., Hodgins J., Matthews I.: A deep learning approach for generalized speech animation. *ACM Trans. Graph. 36*, 4 (July 2017), 93:1–93:11. 1, 2

[TT14] Taylor S., Theobald B.-j.: The effect of speaking rate on audio and visual speech. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (2014), 3037–3041. 2

[Wal82] Walther E. F.: *Lipreading*. Burnham Inc Pub, 1982. 2

[WBLP11] Weise T., Bouaziz S., Li H., Pauly M.: Real-time performance-based facial animation. *ACM Trans. Graph. 30*, 4 (July 2011), 77:1–77:10. 1