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

MOTIVATION

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 on mouth movement.

CONTRIBUTIONS

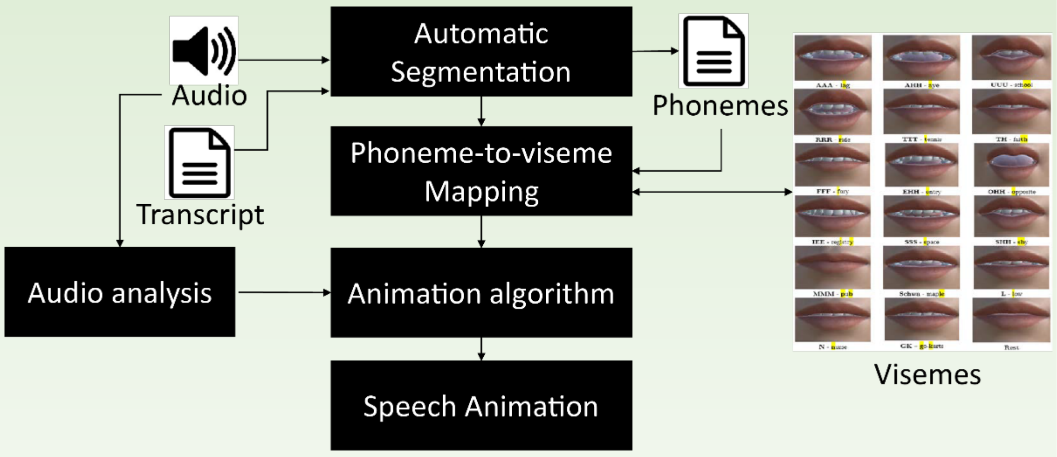
Our contribution is two-fold:

- An expressive speech animation model that takes into account emotional variations in audio;
- A co-articulation model based on dynamic linguistic rules varying among different emotions.

OUR METHOD

- **Input:** An audio file with its corresponding text transcript
- **Output:** Speech animation

The audio file and the 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; the pitch and intensity values are mapped to the weight of the visemes to adjust their intensity. Finally, dynamic co-articulation rules are applied to generate the speech animation.



EXPRESSIVE SPEECH

Assumptions:

- High arousal occurs during intense emotions such as joy or anger, where stronger articulation is observed. Lower arousal is observed with emotions such as sadness or boredom [1].
- Pitch and intensity influence all vowels as well as some of the consonants (plosives and fricatives).
- For vowels, we take the average weight influence of pitch and intensity; whereas for the consonants, we consider the effect of pitch.
- There is a difference between male and female pitch values and that effects the weights of the visemes.

CO-ARTICULATION

Assumptions:

- A viseme starts before its corresponding phoneme, continues rising until it reaches the apex of its phoneme and then starts decaying until it reaches zero.
- Speech rate can strongly influence the way people speak and therefore articulate the lip muscles [2]. With fast speech rates, the duration of and the distance between the phonemes is minimal.
- During fast speech, phonemes can drop or can be substituted.

RESULTS

We compared our method with two commercial speech animation tools by generating 34 videos with various emotions for male and female voices (3 voices per gender).

Emotion	Comparison 1		Comparison 2	
	Ours	**FaceFX**	**Ours**	**RogoDigital**
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
Total	**510**	**374**	**504**	**380**
Total%	**58%**	**42%**	**57%**	**43%**

Conclusion: Our model scores better except the poem which is due to the long duration of the clip with more emotional variations.

Future work: We plan to compare our method with other procedural lip-sync methods (e.g. JALI [3]) and extend to other languages.

REFERENCES

[1] 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.

[2] Taylor S., Theobald B. J.: The effect of speaking rate on audio and visual speech, IEEE ICASSP (2014), 3037-3041.

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

ACKNOWLEDGEMENT & CONTACT

This work is partly funded by the Horizon 2020 RAGE - Realizing an Applied Gaming Eco-system project (Grant no. 644187). For more information, contact z.yumak@uu.nl

Utrecht University

EUROGRAPHICS 2018