Received: 25 April 2019 | Accepted: 29 April 2019

DOI: 10.1002/cav.1892

SPECIAL ISSUE PAPER

WILEY

Audio-driven emotional speech animation for interactive virtual characters

Constantinos Charalambous | Zerrin Yumak | A. Frank van der Stappen

Department of Information and Computing Sciences, Utrecht University, Utrecht, The Netherlands

Correspondence
Zerrin Yumak, Department of Information and Computing Sciences, Utrecht University, 3512 JE Utrecht, The Netherlands.
Email: z.yumak@uu.nl

Funding information
Horizon 2020 RAGE - Realizing an Applied Gaming Eco-system project, Grant/Award Number: 644187

Abstract

We present a procedural audio-driven speech animation method for interactive virtual characters. Given any audio with its respective speech transcript, we automatically generate lip-synchronized speech animation that could drive any three-dimensional virtual character. The realism of the animation is enhanced by studying the emotional features of the audio signal and its effect on mouth movements. We also propose a coarticulation model that takes into account various linguistic rules. The generated animation is configurable by the user by modifying the control parameters, such as viseme types, intensities, and coarticulation curves. We compare our approach against two lip-synchronized speech animation generators. Our results show that our method surpasses them in terms of user preference.

KEYWORDS

audio-driven speech animation, emotional speech, procedural animation

1 | INTRODUCTION

Modern games require a high level of realism to create a bond between the player and the game characters. A high level of realism implies high-quality graphics, a strong game play, and a powerful story mode. However, the communication between the player and the game depends on how the virtual character expresses his/her story: This includes the way he/she moves, acts, and talks. Although there are some games with realistic facial expressions, generating those expressions requires costly hardware and expert animators. Recent work in facial animation produced impressive results, such as the characters in the game Hellblade: Senua's Sacrifice and the MEETMIKE Project.[1] Despite considerable progress in this area, automatically generating lip-synchronized animation stays as an open problem.

Different approaches to lip-synchronized speech animation were developed over the years. Procedural approaches are better in terms of the control of the animation, whereas they might not reach the level of naturalness in performance-capture[2,3] and data-driven approaches.[4,5] Our goal is to develop an audio-driven speech animation method for interactive game characters where the control aspect is of 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 take into account the effect of emotions on the mouth movement. Recently, Edwards et al.[6] have introduced the JALI model to simulate different speech styles controlling the jaw and lip parameters in a two-dimensional viseme space. However, they do not take into account distinct emotional categories and fast/slow speech. Previous works on emotional speech animation learn the emotional profiles of mouth movements from speech data and apply these styles

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

explicitly to modulate the animation either with a procedural[7] or a data-driven approach.[3] Our goal is to simulate the mouth movement during emotions, such as happy, sad, and angry, by controlling viseme intensities with audio features directly and implicitly. Figure 1 shows an example viseme sequence generated with our model (see Section 3.2 for the explanation of the notation in the image caption).

Our contribution is 2-fold: (1) an expressive speech animation model that takes into account emotional variations in audio; (2) a coarticulation model based on dynamic linguistic rules.

Our method has several advantages.

- We show that audio features can be used to create emotion-enabled mouth movements without defining explicit emotion rules.
- The method works with a standard rig for game characters enabled with viseme blendshapes.
- It is configurable with several parameters and can be enhanced with new phonetic and linguistic rules.

In the next section, we mention the related work, followed up with the explanation of our methodology in Section 3. The experiment and results are presented in Section 4. We conclude with limitations and future work in Section 5.

2 | RELATED WORK

Methods for speech animation analyze the underlying mechanisms that drive the facial movements using the acoustic and linguistic features derived from audio and text segments. Typically, phoneme boundaries are extracted from the given sequence either manually or automatically using a speech recognition tool. The phonemes are then mapped to visemes that are their visual counterparts. Multiple phonemes are often mapped into a single viseme, which represents a specific pose of the character's mouth at the apex of the respective phoneme. Recently, it has been shown that there is a many-to-many relationship between phonemes and visemes.[8]

The two main approaches for audio-driven speech animation are procedural and data driven. Data-driven approaches,[8–11] more recently using deep learning,[4,5] find a balance between naturalness versus the control of the animation, but they rely on expensive data collection and processing. Procedural approaches are a better choice in terms of the control of the animation, but they might not reach the level of naturalness in data-driven approaches. Our goal is to develop a configurable procedural approach for emotionally expressive game characters. The animation is amenable to further improvement by animators, by adjusting the visemes' animation curves.

An important aspect in procedural speech animation is coarticulation, that is, how each viseme is effected by the preceding or following phonemes. There are two main approaches: dominance functions[12,13] and rule-based coarticulation.[6,14] The former determines the weight of each phoneme against their neighboring ones and their influence on the respective facial control parameters. However, this method suffers from the inability to provide realistic results especially in representing bilabial and labiodental consonants. To handle these issues, Cosi et al.[15] and King and Parent[16] proposed improved versions of the dominance model. An example of a rule-based method is that by Pelachaud et al.,[14] who introduced forward and backward coarticulation rules. Such rules imply that visemes should influence the shape of their neighboring ones, depending on their type and distance. Xu et al.[17] defined configurable animation curves for visemes by using diphone coarticulation. More recently, Edwards et al.[6] introduced categorical rules: constraints, conventions, and habits. Each category defines an explicit set of linguistic rules that should be applied to ensure the correctness of the final animation (i.e., lip-heavy visemes start early and end late).

Although there are various approaches to procedural speech animation, there are only a few that take into account the effect of emotions on the mouth movement. Bevacqua and Pelachaud[7] captured and analyzed data from a speaker in vowel–consonant–vowel segments in varying emotional categories. This resulted in the production of so-called speech curves per viseme in varying coarticulation contexts and with different emotions. Albrecht et al.[18] generated facial animations by mapping audio features and facial expressions to a two-dimensional arousal and valence space. Taylor et al.[19]

analyzed the similarity between phoneme and viseme sequences in fast and slow speech. Edwards et al.[6] introduced the JALI model to simulate different speech styles. They created the JALI viseme field controlled by jaw (JA) and lip (LI) parameters, which are effected by the pitch and intensity values in the audio. High frequency was related with stronger articulation, hence higher parameter values, whereas lower frequency was related with weaker articulation, hence lower parameter values.

In our work, we correlate the audio features with the weight of the respective visemes to capture the effect of distinct emotional categories, such as happy, sad, and angry, and take into account the frequency variations among males and females. We also apply coarticulation rules using dynamic onset/offset values and speech rate.

3 | OVERVIEW

In this section, we present our approach for audio-driven speech animation. Figure 2 shows the overall pipeline. An audio file with its corresponding transcript is given as input to a phoneme extraction tool, and the phonemes given as output are mapped to their visual counterparts. The audio is analyzed further, and frequency and intensity values are used to adjust the intensity of the visemes. Viseme curves together with the coarticulation rules lead to the final speech animation.

3.1 | Automatic segmentation

There are several tools that can be used to extract phonemes from raw audio, such as Sphinx[20] and the hidden Markov model toolkit.[21] However, they suffer from inaccurate and sometimes misaligned phoneme sequence intervals. Hence, we proceeded to alternative solutions and specifically to forced alignment methods. Forced alignment is used as a phoneme mapping tool. A text transcript is given as input together with the audio. The text transcript is converted into phonemes, and the audio signal is aligned with the phonemes by finding the best fitting time intervals. In our work, we used the Munich Automatic Segmentation System,[22] which had the most accurate results. It is important to mention that phoneme segmentation can still suffer from the general inaccuracy defects of such systems due to variations in accent or noise in the audio signal.

3.2 | Phonemes to visemes

Our work is based on the viseme list introduced in the work of Edwards et al.[6] with slight variations. Similar to the work of Edwards et al.,[6] we apply a many-to-one mapping between phonemes and visemes. For example, phonemes such as *m*, *b*, and *p* have the same visual representation. Visemes *GK*, *L*, and *N* are visualized using tongue-only manipulation, as the mouth movement during these visemes is minimal. Similar to the work of Edwards et al.,[6] the viseme *TTH* is divided into two separate visemes: *TTT* and *TH*. An example is for the word *theta*. Two of the phonemes of this word, *t* and *th*, are mapped with the same viseme *TTH*. However, both phonemes sound and look differently, as the latter requires the teeth to be in contact with the tongue, whereas the former does not. Although most phonemes are correlated with a single letter, there exist a few that are correlated with two alphabet letters. These phonemes are known as diphthongs or diphones. Current speech animation techniques map these diphones to a single viseme. During our experiments, we observed that

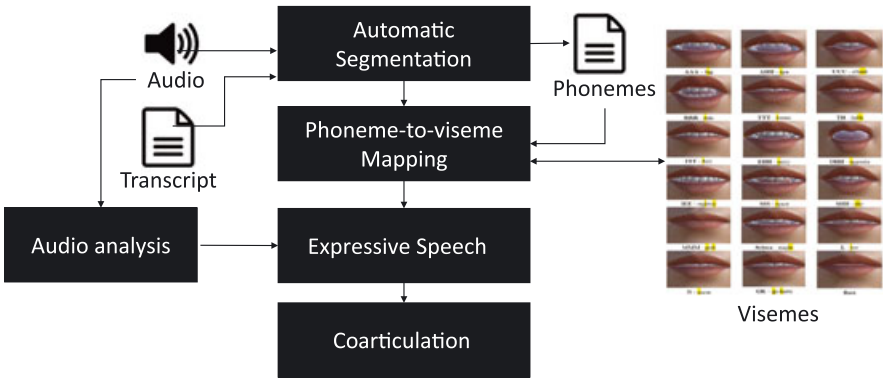

FIGURE 2 Overall pipeline

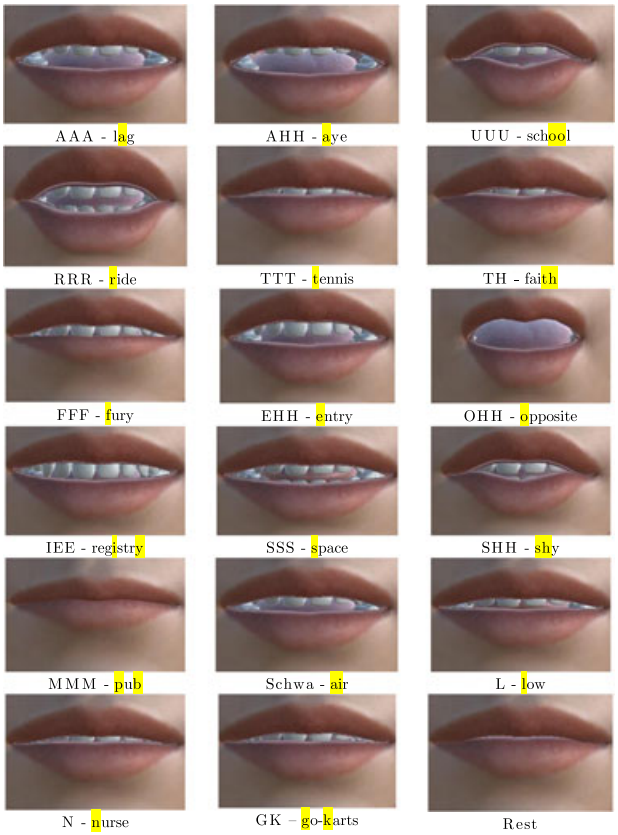

FIGURE 3 Visemes used in our approach

mapping these phonemes to a pair of visemes instead of a single viseme led to more natural animation (see Figure 3 for the full list of visemes).

3.3 | Expressive speech

Pitch and intensity influence all vowels and several consonants during articulation. First, regarding vowels, several studies showed that high mean fundamental frequency, which is the metric for pitch, and high mean intensity (in decibels) are correlated with high arousal.[18,23] High arousal usually appears on several intense or exaggerated emotions such as joy or anger, emotions in which emphasis and stronger articulation occur. On the other hand, lower arousal is correlated with lower fundamental frequency and lower intensity, and it appears in milder or low-intensity emotions such as sadness or boredom. Similarly, lower frequency and intensity imply weaker articulation. Furthermore, frequency has an impact on the articulation of some consonants. This impact mostly appears on specific consonant categories, fricatives and plosives.[6] Fricatives are articulated by pushing the air past the teeth or tongue. Thus, high mean frequencies can be observed during stronger articulations.[24] Plosives are articulated by stopping the airflow by lips or tongue. Since they are heavily articulated, high mean frequencies are observed. Likewise, lower mean frequencies yield weaker articulation. It is important to mention that for the rest of the consonants, we do not account pitch and intensity.

Frequency values differ between male and female. Male frequency varies from 85 to 180 Hz, whereas female frequency varies from 165 to 255 Hz.[25] Similarly, Traunmüller and Eriksson[26] stated that the average male frequency tends to be at approximately 120 Hz, whereas the average female frequency tends to be around at 210 Hz. In our model, we used the frequency ranges as stated in the works of Titze and Martin[25] and Traunmüller and Eriksson.[26]

To calculate pitch and intensity from audio, we used the open-source audio feature extractor openSmile.[27] Pitch is calculated in hertz, using the fundamental frequency metric *F0*, and intensity is calculated using the root-mean-square energy value over a window of 1 ms. We calculated the mean frequency and intensity values for each individual phoneme and mapped these values with the mean blendshape weights. This implies that if the weight of the blendshape of a viseme reaches a value higher than the mean blendshape weight, we will have the phenomenon of hyper-articulation and, in general, stronger articulation. Similarly, a lower weight will induce the phenomenon of hypo-articulation and, therefore, weaker articulation.

The value of a blendshape weight ranges between 0 and 1, with 0 being mouth closed and 1 being the viseme's full morph target. We calculate the weights as in Equations (1) and (2). While W_{F} denotes the weight caused by changes in frequency, W_{I} denotes the weight that is effected by the changes in intensity. As mentioned above, for the vowels, we consider both frequency and intensity and, thus, take the average of W_{F} and W_{I}. For the consonants (plosives and fricatives), we only use the pitch-influenced weight W_{F}.

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

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

The equations imply a linear mapping between a phoneme's intensity and frequency in the given audio clip and the blendshape weights of the corresponding visemes. Denoted with F_{p} and I_{p} are the frequency and intensity values for the respective phoneme p. $\overline{F}$ and $\overline{I}$ are calculated by measuring the frequency and intensity of each phoneme, summing them up and dividing by the number of phonemes in the given audio clip. $\overline{W}_{\mathrm{p}}$ is the mean blendshape weight. $F_{\min}$, $I_{\min}$ and $F_{\max}$, $I_{\max}$ are the minimum and maximum frequency and intensity values for the whole audio clip. Similarly, $W_{\min}$ and $W_{\max}$ represent the minimum and maximum blendshape weights.

3.4 | Coarticulation

Another important aspect of speech is coarticulation. Recall that coarticulation is about how each viseme is effected by the preceding or following phonemes. There are no agreed-upon uniform rules of coarticulation that are defined by linguistic theory as it depends on the context of the visemes. Therefore, existing speech animation approaches deal with coarticulation by experimenting with different rules to increase the level of naturalness.

For smooth blending, we represent the visemes using a quadratic curve with an ascent and a descent part as shown in Equation (3). The weight of a viseme begins with zero and gradually grows until it reaches its maximum value, matching the phoneme's apex. The weight gradually falls starting from the maximum value to zero for the phoneme's remaining time. $W_{P_i}(t_c)$ is the viseme's weight for the phoneme P_i at the current time t_c. D_{V_i} denotes the time interval the animation of a viseme starts and ends. Coefficients a and b are calculated by dividing t_c with the percentage of the ascent and the descent part. This percentage is defined as 75% based on the study of human articulatory muscles.[28]

$$\begin{aligned} \text{EaseIn: } & W_{P_i}(t_c) = a^2 * W_{\max} \\ \text{EaseOut: } & W_{P_i}(t_c) = b * (2 - b) * W_{\max} \end{aligned} \tag{3}$$

$$\begin{aligned} a &= \frac{t_c}{D_{V_i} * 0.75} \\ b &= \frac{t_c}{D_{V_i} * 0.25} \end{aligned} \tag{4}$$

Figure 4 shows two consecutive visemes. According to linguistic studies, the mouth starts to move before a phoneme is vocalized and ends after it is pronounced.[29] Thus, an animation of a viseme starts slightly before its phoneme and ends slightly after it (t_{onset} and t_{offset}). The viseme reaches its maximum t_{apex} by the time the phoneme is completed 75%, as stated above. D_{P_i} and D_{P_j} are the durations of the two consecutive phonemes, as given by the forced alignment step, and D_{V_i} and D_{V_j} are the durations of the corresponding visemes.

According to the sensory–motor studies of speech movement, the onset/offset value is defined to be 120 ms.[29] Edwards et al.[6] used 120 ms as the default onset and offset value for most phonemes and 150 ms for the ones that had the effect of lip protrusion as well as defined other contextual varying onset times. By using a static interval of 120 ms as in the work of Bailly,[29] a major problem arises. On audio with a fast speech rate, the duration and distance between phonemes are small, causing several phonemes to be overlapping each other and, thus, creating misleading and unnatural speech animation.

To counter this problem, we make both the onset and the offset interval as variables so that they can vary depending on the speech rate and the type of the phoneme. We added an influence parameter c that is multiplied with t_{onset} and

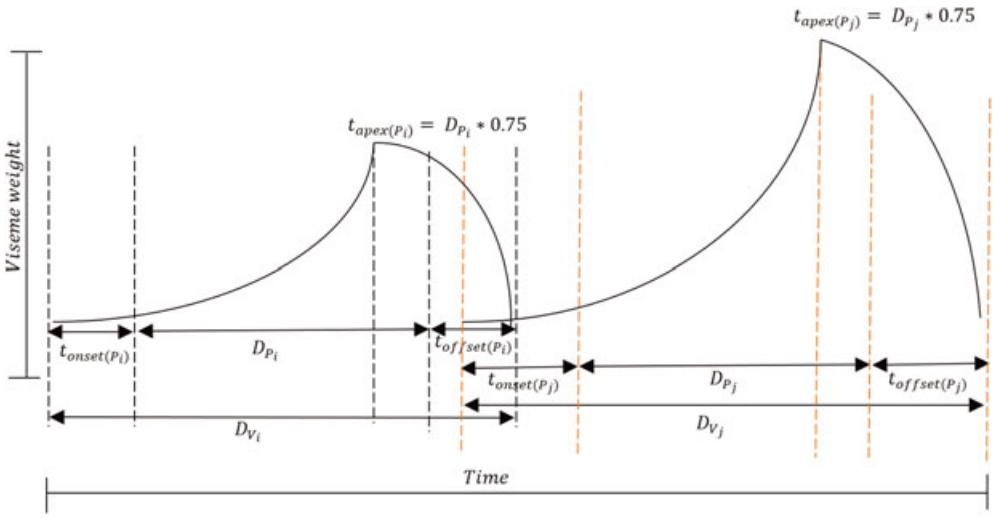

FIGURE 4 Coarticulation for two neighboring phonemes

t_{offset}. Here, c is a weight value varying between 0 and 1 that defines the degree of influence of a phoneme type over another phoneme type. There are four different phoneme influence coefficients: vowel over vowel, vowel over consonant, consonant over vowel, and consonant over consonant. We consider these parameters for the dynamic onset/offset values only on audio with a fast speech rate. For slower speech rates, onset/offset is used as defined by Bailly[29] and Edwards et al.[6]

Figure 5 shows three different scenarios for the emotion type *angry*, where the value of c is set to 0.1, 0.5, and 1, respectively. The audio transcript is *"kids are talking by the door"* taken from RAVDESS, the Ryerson Audio-Visual Database of Emotional Speech and Song.[30] The sentence is converted into SAMPA notation: "k I d z A: t O: k I N b aI D @ d O:". Subsequently, phoneme-to-viseme mapping turns the sentence into "GK IEE T SSS AHH T OHH GK IEE GK MMM AHH IEE TH Schwa T OHH RRR" as defined in Section 3.2. The variations in the top values of the viseme curves are based on the changes in their weight calculated based on pitch and intensity. As the influence parameter c increases, the consecutive phonemes start to overlap with each other more. When the influence parameter is set to $c = 0.1$, most of the visemes do not overlap with each other since their durations are minimized. It means that these visemes will not influence their neighboring ones, and therefore, we will have the phenomenon of fast mumbling due to their minimized duration. Subsequently, by setting the parameter to $c = 0.5$, the neighboring influence is increased in a way that some visemes start to overlap with each other. Finally, when we use $c = 1$ as the value of the influence parameter, we have no influence on the t_{onset} and t_{offset} values at all; therefore, the raw durations obtained from the phoneme extraction tool and the predefined values for t_{onset} and t_{offset} are used. As a result, we observe more overlapped visemes than in the previous cases. For our experiment, same values (vowel over vowel = 0.8, vowel over consonant = 0.4, consonant over consonant = 0.2, and consonant over vowel = 0.4) are used for all the animation clips generated, as these values are taken as emotion independent but rather related with the relationship between phoneme types. They are derived heuristically by observation and based on the above guidelines. Further studies are required to see if these values change according to different emotions.

In addition to the coarticulation model described above, explicit rules of constraints, conventions, and habits are applied, as inspired by the work of Edwards et al.[6] Each category includes various rules that ensure correctness and smoothness for phonemic transition. Constraints include visual rules such as the closure during bilabials (*b*, *p*, and *m*) or contact between the top teeth and the bottom lip during labiodentals (*f* and *v*). Conventions include rules such as strong articulation on lexically stressed vowels (e.g., phoneme *i* in *fire*) or mouth opening during pauses. Finally, habits include rules in which the appearance of each viseme is influenced by neighboring visemes, for example, the union of duplicated visemes (e.g., *pop band*, where *p* and *b* are articulated into a long viseme *MMM*) or neighboring viseme influence during tongue-only visemes (*l*, *n*, *t*, and *d*).

Apart from specifying these explicit coarticulation rules, phonemes that contain more than one letter, known as diphones, are manipulated differently. Since diphones are treated as single phonemes at the forced alignment step, they have a single time duration. We divide their duration into two equal parts and used the phoneme prominence model described above to decide on the onset/offset values. It is known that speech rate can strongly influence the way people speak and, therefore, articulate their lip muscles. We mentioned earlier that by introducing the onset/offset variable, the problem of speech rate is resolved to some extent. On top of that, based on the findings of Taylor et al.,[19] we added extra

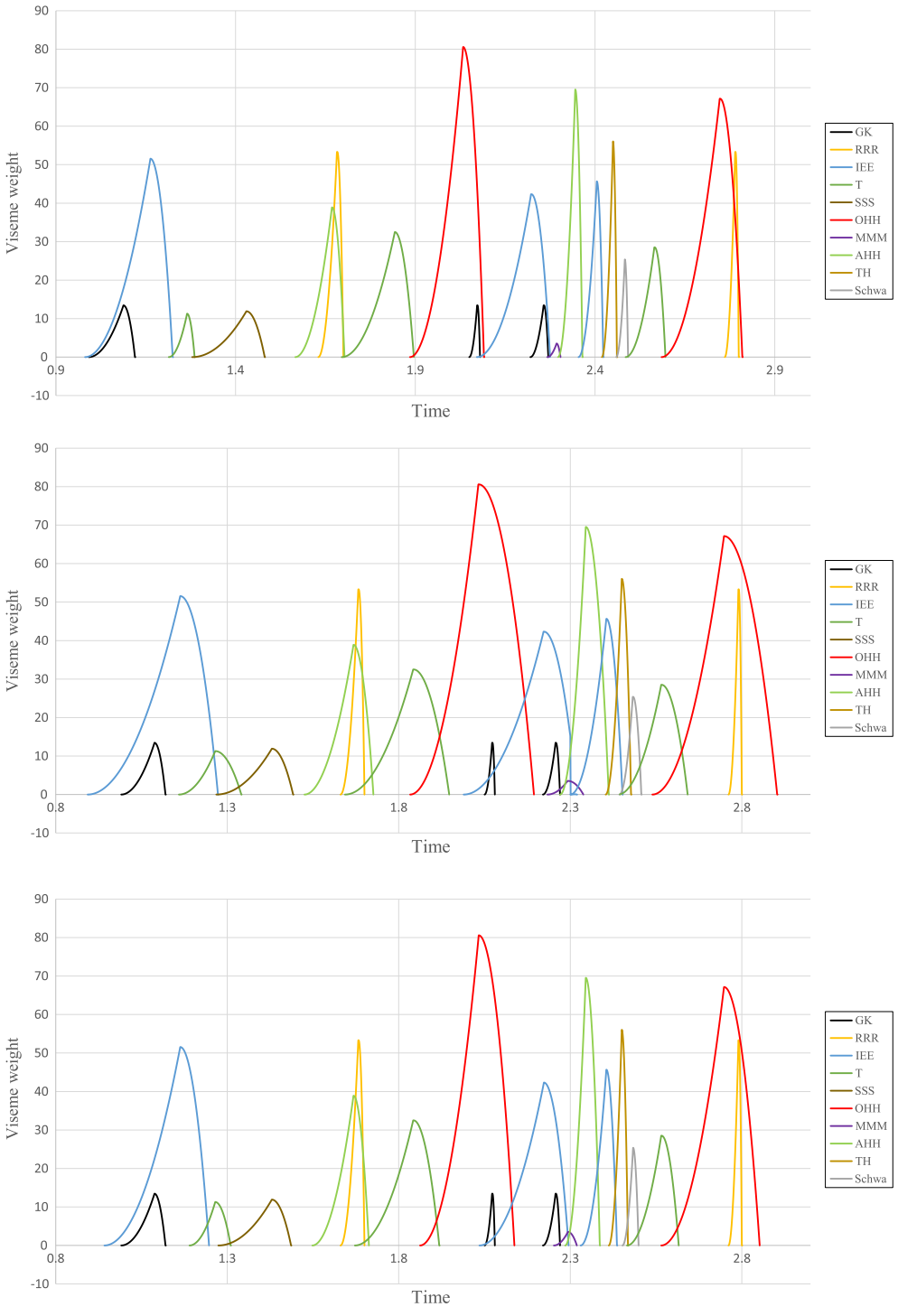

FIGURE 5 Animation curves for influence parameters: $c = 0.1$ (top), $c = 0.5$ (middle), and $c = 1$ (bottom)

rules related to phoneme substitution or deletion. In case of fast speech, several consonants exceeding the threshold are dropped from speech (*h* and *t*), and vowels exceeding the threshold are substituted with the viseme *Schwa*.[31]

4 | EXPERIMENT AND RESULTS

We conducted a user experiment by comparing our method against two speech animation generators: FaceFX[32] and Rogo Digital's LipSync.[33] They are both based on procedural speech animation but do not take into account emotional

FIGURE 6 Male and female virtual characters

variations. External batch scripts were developed for processing the phonemes and audio features based on the Munich Automatic Segmentation System and openSMILE. Our model is implemented as a plug-in to the Unity 3D game engine. The 3D models and blendshapes were created with Daz 3D Studio. Figure 6 shows the male and female 3D models used in our experiments.

For generating the videos, we used the RAVDESS data set,[30] which contains emotional variations of two sentences: *"kids are talking by the door"* and *"dogs are sitting by the door"*. We also extracted longer audio files as poems from the public audiobook library LibriVox. We generated 34 videos representing various emotional states (happy, sad, angry, song, and poem) for male/female voices (three different voices per gender) as seen below.

Six different voices (*FemaleV0, FemaleV1, FemaleV2, MaleV0, MaleV1, MaleV2*)
Five emotional states (*happy, sad, angry, song, poem*)
Four types of transcripts (*"kids are talking by the door", "dogs are sitting by the door", Spring Cowardice, To A Traveler*): the first two transcripts are from the RAVDESS data set, whereas the latter two are poems from LibriVox

We had 26 participants in total, of which 16 are male and 10 are female. The users compared two videos each time, our method versus FaceFX and our method versus Rogo Digital, and selected the one that looks most natural to them. The videos generated for the questionnaire can be viewed online.[34] The results can be seen in Tables 1 and 2.

Our method had a higher user preference over both FaceFX and Rogo Digital, that is, 58% and 57%, respectively. Regarding emotional variations (Table 1), our method scores better in the happy, sad, angry, and song categories. However, the effect was not the same for the poem category. We speculate that that is due to the complex emotional variety in the poem with alternating intonations and longer duration. Our method takes into account shorter audio clips and calculates the mean intensity and frequency values from the whole audio clip. This problem can be encountered by calculating these

TABLE 1 Results (emotional variations)

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
Total	**510**	**374**	**504**	**380**
Total%	**58%**	**42%**	**57%**	**43%**

TABLE 2 Results (voice type)

Voice	Comparison 1		Comparison 2	
	Ours	FaceFX	Ours	Rogo
FemaleV0	108	100	97	111
FemaleV1	140	68	123	85
FemaleV2	11	15	11	15
MaleV0	120	88	143	65
MaleV1	119	89	121	87
MaleV2	12	14	9	17
Total	**510**	**374**	**504**	**380**
Total%	**58%**	**42%**	**57%**	**43%**

values over a sliding window of shorter durations rather than based on fixed-length audio clips. Regarding voice types with male and female voices, the preference is mostly in our favor except for *FemaleV2* and *MaleV2* in the case of the poem again.

During our experiments, we observed that the minimum and maximum frequency values for male and female had a big impact on the final animation. When the default values for the minimum and maximum frequencies introduced in Section 3.3 were not appropriate, we extracted them from the data set we have. We also experimented with swapping the female range with the male range to see whether we see a difference in the final animation. What we noticed is that when we swapped the frequency range from female to male, we had the phenomenon of hypo-articulation, as the average male frequency tends to be significantly lower than the female one. The opposite was observed when we swapped the frequency range from male to female. Since the male frequency range is lower than the average female frequency, we had the phenomenon of hyper-articulation, as visemes tend to reach their maximum weight.

5 | CONCLUSION AND FUTURE WORK

Our approach succeeded in providing mouth animation that was accurate in terms of synchronization with audio and with the ability to express emotional variations. We compared our method against two procedural methods. The results show that our method produces more natural animation in most of the cases. However, there are also limitations and directions for future work.

The results were mainly based on a subjective user study. A quantitative error metric can be added to compare the lip-synchronized motion generated with our method to the ground-truth animation. We also intend to compare our method against other recent audio-driven procedural and data-driven speech animation approaches, conducting another user study with also statistical significance.

Mapping frequency and intensity features from audio with viseme weights was a good starting point toward generating expressive speech. However, further acoustic features and their influence on the visemes should be analyzed. We need to test with other emotional categories to see how the model handles these cases and which improvements can be made. That requires extensive analysis of emotional curves of visemes in various contexts and the construction of an emotional viseme space that can be linked to audio features. Furthermore, although we were able to generate variations between fast and slow speech animation, the results for the faster animation are found to be less natural, which requires further investigation.

The results were heavily dependent on the phoneme segmentation tool. The system requires both audio and its respective transcript to extract the phoneme intervals, which are important for accurate synchronization. However, for a fully automatic pipeline, we need to surpass the audio's transcript and purely extract phoneme information by using only the audio. There exist several tools that already implemented this procedure, but with low accuracy. Finding the perfect tool for phoneme extraction still remains a challenge.

Finally, we need to add more facial parameters that are influenced by speech and not be restricted only on mouth movement: Cheeks, eyes, and head are parameters that are heavily influenced by speech. Moreover, the model should be tested with other languages.

ACKNOWLEDGEMENTS

This work was partly supported by the Horizon 2020 RAGE (Realizing an Applied Gaming Ecosystem) Project under Grant 644187. We would like to thank the participants for attending the user study.

ORCID

Zerrin Yumak https://orcid.org/0000-0002-0028-5806

REFERENCES

1. Seymour M, Evans C, Libreri K. Meet mike: epic avatars. Paper presented at: ACM SIGGRAPH 2017 VR Village; 2017 July 30–Aug 3; Los Angeles, CA. New York, NY: ACM; 2017.
2. Weise T, Bouaziz S, Li H, Pauly M. Realtime performance-based facial animation. ACM Trans Graph. 2011;30(4):77:1–77:10.

3. Cao C, Wu H, Weng Y, Shao T, Zhou K. Real-time facial animation with image-based dynamic avatars. ACM Trans Graph. 2016;35(4):126:1–126:12.
4. 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. 2017;36(4):94:1–94:12.
5. Taylor S, Kim T, Yue Y, et al. A deep learning approach for generalized speech animation. ACM Trans Graph. 2017;36(4):93:1–93:11. http://doi.acm.org/10.1145/3072959.3073699
6. Edwards P, Landreth C, Fiume E, Singh K. JALI: an animator-centric viseme model for expressive lip synchronization. ACM Trans Graph. 2016;35(4):127:1–127:11.
7. Bevacqua E, Pelachaud C. Expressive audio-visual speech. Comput Animat Virtual Worlds. 2004;15(3-4):297–304.
8. Taylor SL, Mahler M, Theobald B-J, Matthews I. Dynamic units of visual speech. Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation; 2012 July 29–31; Lausanne, Switzerland. Eurographics Association; Goslar, Germany; 2012. p. 275–284. http://dl.acm.org/citation.cfm?id=2422356.2422395
9. Bregler C, Covell M, Slaney M. Video rewrite: driving visual speech with audio. Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques; 1997 Aug 3; Los Angeles, CA. New York, NY: ACM Press/Addison-Wesley Publishing Co.; 1997. p. 353–360. https://doi.org/10.1145/258734.258880
10. Deng Z, Neumann U, Lewis JP, Kim T-Y, Bulut M, Narayanan S. Expressive facial animation synthesis by learning speech coarticulation and expression spaces. IEEE Trans Vis Comput Graph. 2006;12(6):1523–1534. https://doi.org/10.1109/TVCG.2006.90
11. Kim T, Yue Y, Taylor S, Matthews I. A decision tree framework for spatiotemporal sequence prediction. Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2015 Aug 10–13; Sydney, Australia. New York, NY: ACM; 2015. p. 577–586. http://doi.acm.org/10.1145/2783258.2783356
12. Cohen MM, Massaro DW. Modeling coarticulation in synthetic visual speech. In: Thalmann NM, Thalmann D, editors. Models and techniques in computer animation. Tokyo, Japan: Springer; 1993. p. 139–156.
13. Massaro DW, Cohen MM, Tabain M, Beskow J, Clark R. Animated speech: research progress and applications. In: Bailly G, Perrier P, Vatikiotis-Bateson E, editors. Audiovisual speech processing. Cambridge, UK: Cambridge University Press; 2012. p. 309–345.
14. Pelachaud C, Badler NI, Steedman M. Generating facial expressions for speech. Cognitive Science. 1996;20(1):1–46. http://www.sciencedirect.com/science/article/pii/S0364021399800019
15. Cosi P, Caldognetto EM, Perin G, Zmarich C. Labial coarticulation modeling for realistic facial animation. Proceedings of the Fourth IEEE International Conference on Multimodal Interfaces; 2002 Oct 14–16. Pittsburgh, PA. Washington, DC: IEEE Computer Society; 2002. p. 505–510. https://doi.org/10.1109/ICMI.2002.1167047
16. King SA, Parent RE. Creating speech-synchronized animation. IEEE Trans Vis Comput Graph. 2005;11(3):341–352. https://doi.org/10.1109/TVCG.2005.43
17. Xu Y, Feng AW, Marsella S, Shapiro A. A practical and configurable lip sync method for games. Proceedings of Motion on Games; 2013 Nov 6–8; Dublin 2, Ireland. New York, NY: ACM; 2013. p. 131–140. http://doi.acm.org/10.1145/2522628.2522904
18. Albrecht I, Schröder M, Haber J, Seidel H-P. Mixed feelings: expression of non-basic emotions in a muscle-based talking head. Virtual Reality. 2005;8(4):201–212. https://doi.org/10.1007/s10055-005-0153-5
19. Taylor S, Theobald B, Matthews I. The effect of speaking rate on audio and visual speech. Paper presented at: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP); 2014 May 4–9; Florence, Italy. Piscataway, NJ: IEEE; p. 3037–3041.
20. Lee K, Hon H, Reddy R. An overview of the sphinx speech recognition system. IEEE Trans Acoust Speech Signal Process. 1990;38(1):35–45.
21. Young SJ, Young S. The HTK hidden Markov model toolkit: design and philosophy. Cambridge, UK: Department of Engineering, University of Cambridge; 1993.
22. Schiel F, Draxler C, Harrington J. Phonemic segmentation and labelling using the MAUS technique. Paper presented at: Workshop New Tools and Methods for Very-Large-Scale Phonetics Research; 2011 Jan 28-31; Philadelphia, PA.
23. Banse R, Scherer KR. Acoustic profiles in vocal emotion expression. J Pers Soc Psychol. 1996;70(3):614–36.
24. Maniwa K, Jongman A, Wade T. Acoustic characteristics of clearly spoken English fricatives. J Acoust Soc Am. 2009;125(6):3962–3973.
25. Titze IR, Martin DW. Principles of voice production. J Acoust Soc Am. 1998;104(3):1148–1148.
26. Traunmüller H, Eriksson A. The frequency range of the voice fundamental in the speech of male and female adults. Stockholm, Sweden: Department of Linguistics, University of Stockholm; 1994. Technical report.
27. Eyben F, Weninger F, Gross F, Schuller B. Recent developments in openSmile, the Munich open-source multimedia feature extractor. Proceedings of the 21st ACM International Conference on Multimedia; 2013 Oct 21–21; Barcelona, Spain. New York, NY: ACM; 2013. p. 835–838. http://doi.acm.org/10.1145/2502081.2502224
28. Ito T, Murano EZ, Gomi H. Fast force-generation dynamics of human articulatory muscles neural control of movement. J Appl Physiol. 2004;96:2318–2324.
29. Bailly G. Learning to speak. Sensori-motor control of speech movements. Speech Communication. 1997;22(2):251–267. http://www.sciencedirect.com/science/article/pii/S0167639397000253
30. Livingstone SR, Russo FA. The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): a dynamic, multimodal set of facial and vocal expressions in North American English. PloS One; 2018;13(5):e0196391.
31. Lindblom B. Spectographic study of vowel reduction. J Acoust Soc Am 1963;35(11):1773–1781.
32. Maestri G. FaceFX studio 2010. Comput Graph World. 2011;34(4):44–45.

33. Rogo Digital. Rogo Digital LipSync. Available from: https://lipsync.rogodigital.com.
34. Charalambous C. Audio-driven speech animation for virtual characters - questionnaire. Available from: https://questlipsync.herokuapp.com.

AUTHOR BIOGRAPHIES

Constantinos Charalambous has an M.Sc. degree in game and media technology from Utrecht University, Utrecht, The Netherlands, and a B.Sc. degree in computer science from the University of Cyprus, Nicosia, Cyprus. He is currently a Full-Stack Developer at Dynamic Works. His interest includes web and game programming.

Zerrin Yumak is an Assistant Professor of computer science at Utrecht University, Utrecht, The Netherlands. She holds a Ph.D. degree from MIRALab, University of Geneva, Geneva, Switzerland, and has several years of postdoctoral experience at the Swiss Federal Institute of Technology, Lausanne, Switzerland, and Nanyang Technological University, Singapore. Her research interests include expressive character animation, including face, gaze, and gesture animations, and modeling social behaviors using techniques from computer graphics, artificial intelligence, robotics, and human–machine interaction.

A. Frank van der Stappen is an Associate Professor of computing science at Utrecht University, Utrecht, The Netherlands, and an IEEE Fellow. He holds a Ph.D. degree from Utrecht University and PD.Eng. and M.Sc. degrees from Eindhoven University of Technology, Eindhoven, The Netherlands. His research interests include algorithms for robotics and industrial automation, character animation, simulation, three-dimensional modeling, and computational geometry. He is the author of well over 110 papers in major journals and conferences. He has served on various editorial boards and program committees.

How to cite this article: Charalambous C, Yumak Z, van der Stappen AF. Audio-driven emotional speech animation for interactive virtual characters. *Comput Anim Virtual Worlds*. 2019;30:e1892. https://doi.org/10.1002/cav.1892