A Novel Lip Synchronization Approach for Games and Virtual Environments

Jan Krejsa
Masaryk University
Brno, Czech Republic
433582@mail.muni.cz

Fotis Liarokapis
CYENS – Centre of Excellence
Nicosia, Cyprus
f.liarokapis@cyens.org.cy

***Abstract*—This paper presents an algorithm for the offline generation of lip-sync animation. It redefines visemes as sets of constraints on the facial articulators such as lips, jaw, tongue. The algorithm was comparatively evaluated with 30 healthy participants by presenting a set of phrases delivered verbally by a virtual character. Each phrase was presented in two versions: once with the traditional lip-sync method, and once with our method. Results confirm that the suggested solution produces more natural animation than standard keyframe interpolation techniques.**

***Index Terms*—virtual reality, computer games, lip-sync animation, user-studies**

I. Introduction

The purpose of *lip synchronization* (or *lip-sync*) is to reinforce the illusion that the voice of an animated character is coming out of the character's mouth. This can help create an emotional connection to the character, and improve the immersion of the viewer in a virtual environment. There are currently two leading techniques in lip synchronization. One is capturing the performance of human actors via motion capture techniques, the other is hiring professional animators to produce the facial animations manually. Both methods deliver high-quality results, but they are time-consuming and expensive. As such, they are not suitable for applications where a large quantity of speech animation is required (e.g. role-playing games). In such applications, it is preferable to generate the majority of speech animations automatically.

Lip animation can also be automatically generated from audio input or a phonetic transcript. However, although significant advances have been made recently [1], [2], the resulting quality is still not on par with the two aforementioned approaches. In practice, the automated approach tends to be used only to generate low-quality baseline animations. Animators then try to refine as many animations as possible, and for the most important scenes, they often use motion capture.

When synthesizing speech animation, the speech audio we want to synchronize to is usually segmented into atomic units of sound, called *phonemes*. For example, the "m" sound in the word "mother" is described by the phoneme /m/. A *viseme*, then, is a group of phonemes, all of which are visually similar to each other, but distinguishable from phonemes that belong to another viseme [3]. Some phonemes are hard to distinguish visually from one another and multiple phonemes may map to the same viseme.

978-1-6654-3886-5/21/$31.00 ©2021 IEEE

This paper proposes a novel solution to improve the quality of auto-generated lip-sync animation. Even despite the lower quality of its results, the automated approach still has an irreplaceable role due to its time- and cost-saving properties. Traditionally, auto-generated lip-sync animation is achieved via keyframe interpolation. First, the speech audio is segmented into short units known as visemes. For each viseme, the traditional approach defines a viseme shape: the shape of the mouth that is required to produce the given sound.

These poses are used as animation keyframes placed at the onset of each viseme, and the final animation is created by interpolating between these keyframes. The visemic context or co-articulation (i.e. how the appearance of the given viseme is influenced by the visemes that come before and after it) is usually disregarded, although there are works that have suggested solutions. Notable methods involve dominance functions [4] or neural networks [5].

In this work, we suggest an alternative solution for modeling co-articulation ("Fig. 1"). We propose to replace each viseme shape with a set of constraints on certain parameters of the face (e.g. mouth openness). Instead of specifying a single facial pose for each viseme, the constraints define a wide range of allowed poses. Since each viseme in our model has constraints for only some of the facial parameters (e.g. mouth width, jaw openness, etc), the parameters that are not constrained by the given viseme are free to be driven by neighboring visemes. Effectively, this models co-articulation and leads to more fluid and natural motion.

Our method was evaluated in a study with 30 participants. Participants were presented with a set of phrases delivered verbally by a virtual character. Each phrase was presented in two versions: once with the traditional lip-sync method, and once with our method. Participants were asked to rate the lip motion for each phrase in several aspects, such as naturalness or perceived quality.

In all measured aspects, our method significantly outperformed the traditional method. On a scale from 0 to 100%, the naturalness of our method was rated at 70% on average, compared to 56% naturalness for the traditional method. In terms of quality, our method was rated at 75%, while the traditional method received a rating of 60%. Overall, the

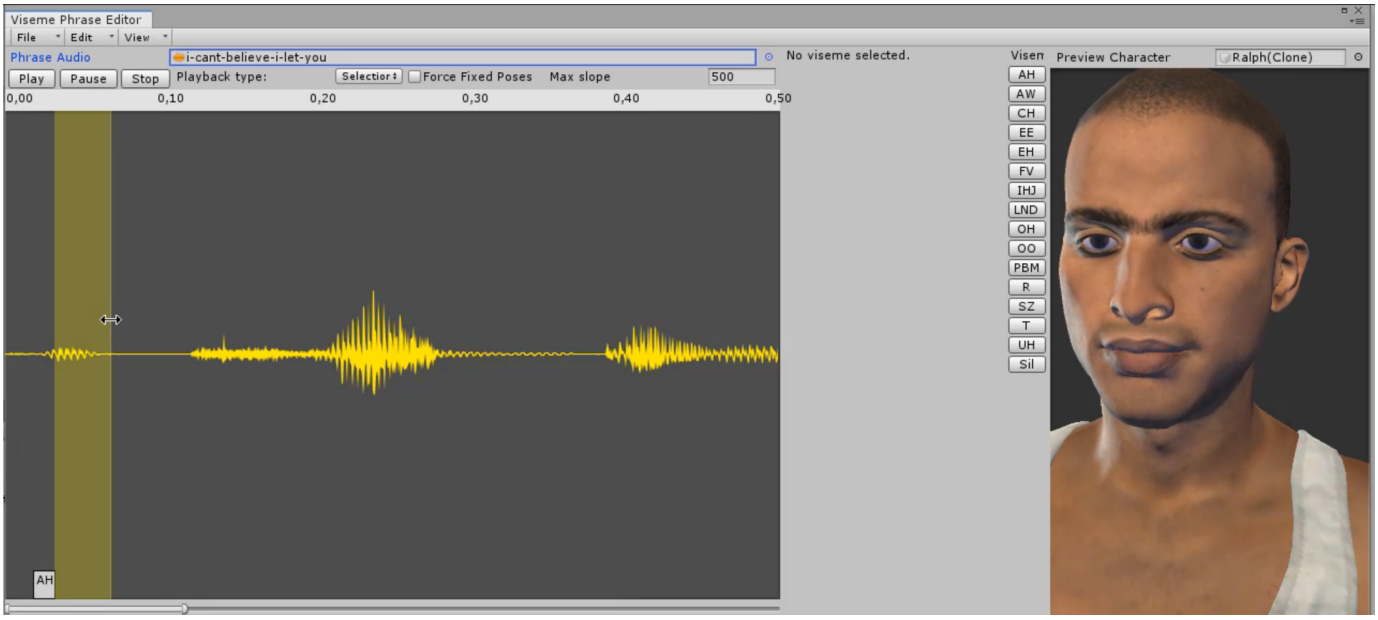


Fig. 1. An example user interface for viseme tagging.

average score of our method was 73%, which is a significant improvement over the 60% rating of the traditional method.

II. Related Work

There are three fundamental approaches to lip synchronization: animation by hand, facial performance capture, and automatic synthesis. Hand-made facial animation has the advantage of having full control over the performance. A skilled artist can add more expressiveness, even beyond the limits of what an actor could do. However, animating by hand is time-consuming, and as such, it may not be suitable for large-scale projects that require many hours of facial animation (such as an expansive role-playing game), especially if their budget is limited. This section will focus on the other two approaches (facial performance capture and automatic synthesis).

A. Facial Performance Capture

Performance capture records motion data of a human actor and transfers it onto a virtual character [6]. By recording the performance of an actor, we can capture all the intricacies of their facial movements in great detail. This approach currently achieves the highest-quality results, but it is also the most expensive method. The current industry standard in film is to track markers on the face using a helmet-mounted camera (HMC) rig [7]. Tracking precision can be an issue, requiring an animator to clean up the recorded data manually, which can be quite laborious. The resulting animations are difficult to edit or refine [8]. The method also generally requires expensive equipment and professional actors.

Since motion capture equipment is so expensive, there have been multiple works that attempt to reduce the required cost. Cao et al. [9] have presented a method that adds high-fidelity facial detail to low-resolution face tracking data. However, although the method produces plausible wrinkles, they are still a mere approximation limited by the training data, and it may deliver unexpected results when presented with a pose outside of its training data set. Ma and Deng [10] suggest a geometry-based method for capturing facial performance that includes wrinkle information, while still retaining low-cost acquisition. Their method can reconstruct high-resolution facial geometry and appearance in real-time and with superior accuracy.

Laine et al. [11] utilized deep learning to reduce the amount of manual clean-up and corrections required to maintain a high-quality output. Their pipeline involves manually cleaning up a small subset of the input video footage to train a deep neural network, which is then used to automatically process the remaining video footage at a very high speed. While the training data must be recorded in stereo, the remaining footage may be monocular. In comparison with previous methods, their method delivers increased accuracy, especially in the mouth and eye regions.

B. Automatic Synthesis

Existing methods of automatic synthesis can be grouped into two categories: (1) data-driven methods and (2) procedural methods based on key-frame interpolation.

1) Data-driven Methods: Data-driven methods of lip-sync animation are generally based on either (a) concatenating short speech animation units sampled from a large data set of motion-captured speech, or (b) sampling from statistical models extracted from motion-captured data [12]–[15].

Taylor et al. [5] proposed to replace viseme shapes with animation units, called dynamic visemes. They are short animation units extracted from a training data set, which can then be rearranged in a different order and blended into each other. This technique produces good results for phrases that are

similar to the training data, but to consistently provide good results for any possible phrase, an exhaustive data set that covers all possible sequences of phonemes would be required.

Deng et al. [16] proposed a method that builds co-articulation models based on motion-capture data, using weight decomposition. Again, as with most other data-driven methods, their method would provide best results with a complete training data set of all possible combinations of phonemes. Deena et al. [17], [18] presented an approach based on a switching state model. One disadvantage is that they require the entire training data set to be manually phoneme-tagged.

Asadiabadi et al. [19] used deep learning to train a speaker-independent model that can reproduce the emotions of an actor. In another approach, Long Short-Term Memory (LSTM) networks were used to improve the mapping from phoneme sequences to their dynamic visemes [20], [21]. Karras et al. [2] presented a deep-learning-based method that generates lip-sync animation in real time, based on real-time audio input. Their method derives speech curves directly from audio, reducing the error that is otherwise accumulated over multiple steps. Zhou et al. [8] contributed another solution for automatic real-time lip-synchronization from audio, using LSTM networks. Their method produces output that is more suitable for an animator-centric workflow.

2) Procedural Methods: The traditional way of audio-based procedural synthesis is to analyze input audio to identify visemes, align keyframes (viseme shapes) to these visemes, and then obtain the final animation by interpolating between these keyframes [22]. Uz et al. [23] used physically-based modeling of facial muscles to produce speech animations. The muscles are modeled as forces deforming the polygonal mesh. A specific mouth shape is assigned to each phoneme by setting parameter values that represent the muscles.

Traditional lip-sync methods assign a single specific mouth pose to each viseme. In real speech, however, the appearance of a viseme on the lips (and other articulators) may be influenced by visemes coming before it or after it. For instance, when pronouncing the syllable "moo", the "oo" sound requires the lips to pucker. In anticipation of this, the lips will pucker already during the "m" sound (see "Fig. 2"). This phenomenon is called visual speech co-articulation. The term co-articulation "refers to changes in speech articulation (acoustic or visual) of the current speech segment (phoneme or viseme) due to neighboring speech" [24].

Löfqvist [25] suggested modeling co-articulation using dominance functions. Cohen and Massaro [4] implemented the model proposed by Löfqvist, using negative exponential functions as the dominance functions. Cosi et al. [26] have further improved upon this implementation by adding a resistance coefficient for each dominance (i.e. for each phoneme-articulator pair). King and Parent [27] criticize the choice of negative exponential functions, which are continuous only at the C0 level.

Xu et al. [28] modeled co-articulation by creating animations for transitioning between pairs of phonemes. In another approach, Edwards et al. [1] introduced two parameters: jaw and lip. The values of these two parameters dynamically change depending on the tone of speech detected in the audio. Lazalde et al. [29] proposed a constraint-based approach: each viseme shape is defined as a distribution around an ideal target. An optimization function is used to produce a trajectory that satisfies the constraints given by each distribution, as well as other constraints such as an acceleration/deceleration limit. Our method is based on a similar principle.

Fig. 2. The facial pose produced when pronouncing the phoneme /m/ in the syllable "moo".

III. Implementation

When we speak, we produce various sounds by increasing or decreasing air pressure and constricting the airflow at certain points (throat, tongue, lips, teeth) [30]. To pronounce a phoneme, the face must first satisfy certain conditions. For example, to produce the sound associated with the phoneme /m/, the lips must be closed. Note that this is the only condition for /m/. As long as the lips are closed, the jaw could be open, or the lips may be pursed in anticipation of the next phoneme, and we can still clearly pronounce the /m/. This happens during speech all the time, due to co-articulation.

This is why it is not sufficient to use a pre-defined shape as the facial pose for each viseme. Instead, we suggest defining this pose as a set of constraints on parameters of the speech articulators (e.g. jaw openness, lip pursedness, etc). When lip-syncing, these constraints must be satisfied at the onset of the given viseme in the audio. Since visemes in our model do not have constraints defined for all parameters, the parameters that are not constrained by the given viseme can be driven by nearby visemes. This allows neighboring visemes to mutually influence their appearance, modeling co-articulation.

The facial rig provides controls for the following parameters: Mouth open/close; Mouth wide/narrow; Mouth up/down; Mouth frown; Jaw forward/back; Jaw up/down; Tongue up/down; Lower lip in/out; and Upper lip in/out. Crucially, the facial rig must be constructed such that each pair of opposing articulator parameters (i.e. mouth wide and mouth narrow) must cancel each other out. Next, we define five types of constraints:

- Absolute - e.g. "The mouth must be exactly 0% open"

- Minimum - e.g. “The mouth must be at least 20% open”
- Maximum - e.g. “The mouth cannot be more than 10% open”
- Range - a combination of Minimum and Maximum, e.g. “The mouth must be at least 10%, but not more than 30%, open”
- Relative - e.g. “The mouth goes 5% wider compared to its previous state”

Our viseme definitions are loosely based on viseme definitions suggested by [31]. The algorithm generates animation curves synchronized to a given audio phrase. As input, the algorithm takes a series of timestamps tagged with visemes. These timestamps should be aligned to the phonemes in a speech recording. First, the sequence is split into tracks per articulator parameter. The result of this step is a set of keyframe sequences, each of which corresponds to the motion of one articulator parameter. Each keyframe contains a timestamp and a constraint structure. This step is illustrated in “Fig. 3”, which shows the waveform of the phrase “Nice job!”, with viseme tags below it.

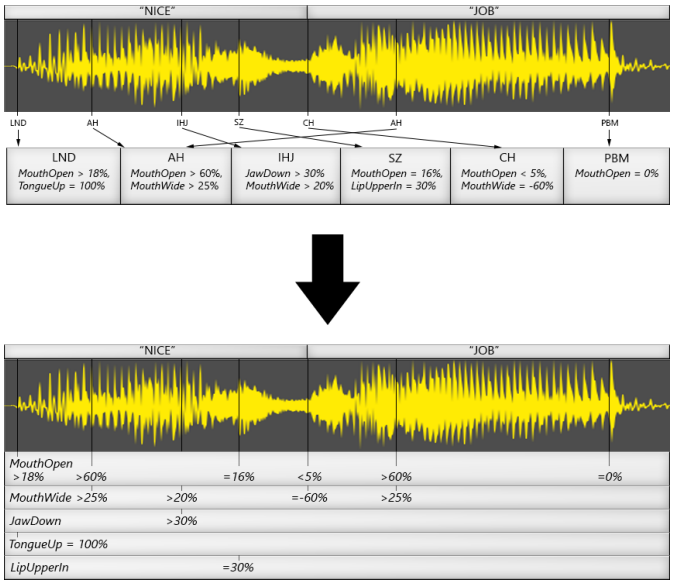


Fig. 3. Visualization of the first step of the algorithm.

Arrows point from each viseme tag to the definition of the corresponding viseme. Below the large arrow, we see the same data transformed into the form of separate keyframe sequences for each parameter. Next, each track is processed separately. The first step when processing one of the tracks is to split the keyframes by their constraint types as shown in “Fig. 4”. At this point, we only have keyframes for the apex of each viseme. We also need keyframes for when the motion starts, i.e. when the mouth starts opening. Wherever there is a long enough pause between keyframes, we insert a keyframe with the value of 0 (i.e. neutral pose). These zero keyframes are added to the list of keyframes with absolute constraints.

For each keyframe with a min/max constraint, we first evaluate the curve in its current state at the given time. If the

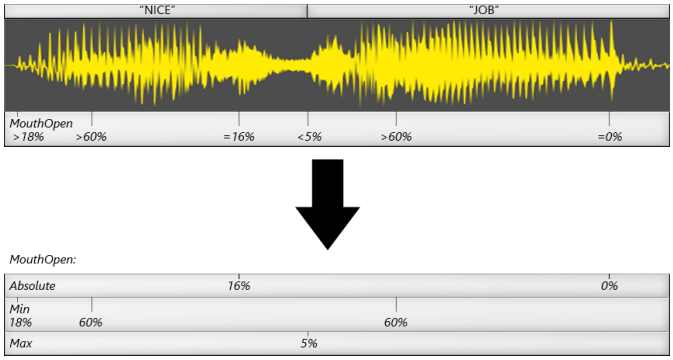


Fig. 4. Visualization of splitting track keyframes by constraint type.

curve already satisfies the constraint, no keyframe is added; otherwise, we add a keyframe with the value of the min/max constraint. Finally, we add relative keyframes by adding their value to the value evaluated from the curve at the given timestamp. The curves obtained by processing all parameter tracks are joined to make up the final animation clip. Due to the physical limitations of the face, there is a limit to the acceleration and deceleration rate of the speech articulators [32]. Our algorithm is capable of generating animations that exceed these limits, in case two visemes occur in quick succession. This gives an appearance of over-articulation. A post-processing pass is applied to correct this.

The user can input the maximum articulator velocity allowed for the given phrase. Note that low values will result in the appearance of mumbling, whereas higher values will result in over-articulation—this may be desirable, depending on the intended style of speech. To enforce the maximum articulator velocity, we process each animation curve separately. In each curve, we examine each pair of successive keyframes. If the slope between the values of these successive keyframes exceeds the maximum velocity, then both keyframes are adjusted to satisfy the velocity constraints. The pseudo-code for this is detailed in “Fig. 5”.

```
PostProcess(AnimationCurve curve)
    Keyframe prevKey ← curve.keys[0];
    for (var i ← 1 to curve.keys.Length)
        Keyframe currentKey ← curve.keys[i];

        float valueDelta ← currentKey.value - prevKey.value;
        float timeDelta ← currentKey.time - prevKey.time;
        float slope ← valueDelta / timeDelta;

        if (Mathf.Abs(slope) > maxSlope)
            float correction ← maxSlope / Mathf.Abs(slope);
            prevKey.value ← correction * prevKey.value;
            currentKey.value ← correction * currentKey.value;
        end

        prevKey ← currentKey;
    end
end
```

Fig. 5. Pseudo-code for post processing the animation curves.

IV. Methodology

A user study was conducted to compare our method to the standard method of blending between static key poses. 30 participants were asked to rate example lip-synchronized phrases in seven aspects:

- Natural ("The lip motion seemed natural.")
- Robotic ("The lip motion seemed robotic.")
- Artificial ("The lip motion seemed artificial.")
- Immersion-Breaking ("The lip motion breaks my immersion.")
- Quality ("I was satisfied with the quality of the lip-sync animation.")
- Temporally synchronized ("The lip animation was synchronized well with the speech audio.")
- Visually correct ("The lip animation matched the speech audio well.")

The phrases used were: "Can you keep a secret?", "Oh yeah, everything's fine." and "I can't believe I let you talk me into this!". Each phrase was presented once with the baseline method and once with our method. The order was randomized for each participant. To aid with immersion, the main part of the experiment was conducted in an immersive virtual reality (VR) environment ("Fig. 6").

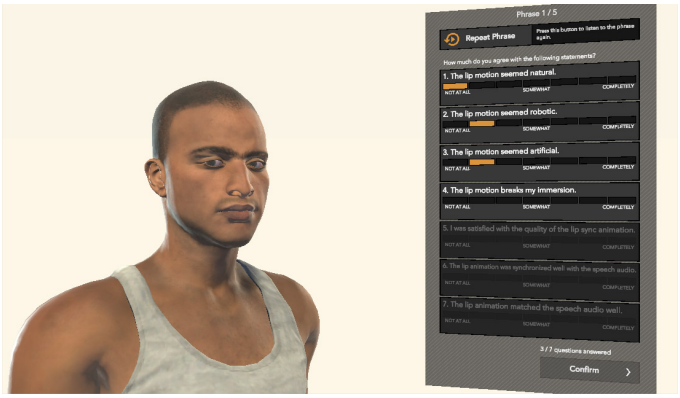


Fig. 6. VR Environment.

A. Participants and Procedure

The sample was made up of 30 test subjects, consisting of 16 male and 14 female participants ("Fig. 7"). Before the start of the experiment, each participant first received an explanation of the task. They proceeded to sign a consent form and fill out their personal information. The users were then instructed on the use of the HTC Vive Pro head-mounted display (HMD). Once they put on the helmet, they were presented with a virtual testing environment in VR. The testing environment contained a virtual character, a button to trigger the playback of the current phrase, and a world-space GUI panel. The GUI panel was used to present questions regarding the current phrase, and to input answers to these questions.

In this environment, participants were allowed to replay any given phrase an unlimited number of times, and to closely examine the lip motion from any angle or distance, taking full advantage of the free movement provided by the VR technology. The questions were presented inside the virtual environment. This was done to avoid having to interrupt the flow by having the user take off the headset repeatedly. Finally, participants were given space to ask questions, express their opinion, and give feedback both on the experiment and on the lipsync animations. Optionally, they could provide written comments on a sheet of paper.

Fig. 7. User participating at the user study.

B. Questionnaires

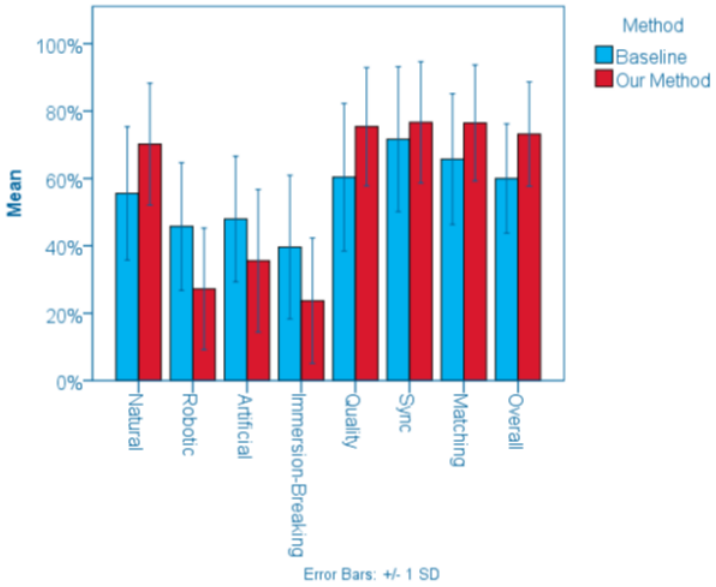
Two different questionnaires were used together with demographics and qualitative comments. The first one is the VR UX questionnaire [33] which focused on the user experience, and the second one is the NASA Task Load Index (TLX) [34] that focused on cognitive demands. The VR UX questionnaire is a compilation of several well-known questionnaires including: presence, engagement, immersion, flow, emotion, judgment, experience consequence, and the technology adoption.

V. Results

A. Questionnaire Results

In the VR portion of the experiment, each participant was presented with 6 lip-synchronized phrases (3 phrases made with the baseline method, 3 with our method). They were asked to rate each phrase in seven aspects. The mean measured values of each aspect for each of the two methods can be seen in "Fig. 8".

The first three aspects (Natural, Robotic, Artificial) were focused on the naturalness of the lip-sync animation. It was presented in three different phrasings to minimize the effects of confusion and misinterpretation. The baseline method was rated as approximately 56% natural on average, whereas our method was rated as 70% natural. As expected, the Robotic



	Baseline		Our Method	
	Mean	SD	Mean	SD
Natural	55.56%	19.79%	70.19%	18.07%
Robotic	45.74%	18.95%	27.22%	18.07%
Artificial	47.96%	18.65%	35.56%	21.17%
Immersion-Breaking	39.63%	21.33%	23.70%	18.62%
Quality	60.37%	21.92%	75.37%	17.55%
Sync	71.67%	21.51%	76.67%	18.01%
Matching	65.74%	19.41%	76.48%	17.24%
Overall Rating	60.00%	16.24%	73.17%	15.48%

Fig. 8. Comparative results.

VR experience aspect	Mean	SD
Presence	83.33%	11.78%
Immersion	55.89%	19.70%
Flow	55.33%	15.56%
Emotion	67.88%	11.81%
Experienced consequences	8.75%	10.56%
Technology adoption	80.00%	9.51%

Fig. 9. VR questionnaire results.

and Artificial ratings closely mirrored the Natural rating for both methods.

The next rated aspect was Immersion-Breaking. The baseline method was rated as approximately 40% immersion-breaking. This could mean that the participants were used to seeing lip animation of poor quality (in comparison with performance-capture animation) and that they can tolerate it. However, our method has significantly improved upon this with an Immersion-Breaking rating of only 24%.

"Fig. 9" shows the mean values of VR experience aspects measured by the questionnaire. the experiment received a low task load rating: 4.96 out of 20. To obtain the overall score, we first multiply each rating by its corresponding weight. Then, we divide the sum of these adjusted ratings by the sum of the

	Mean Rating	SD	Mean Weight	SD
Mental Demand	7.40	4.90	4.07	1.08
Physical Demand	1.10	1.92	0.73	0.83
Temporal Demand	1.37	1.59	1.90	1.12
Performance	3.53	3.16	2.87	1.14
Effort	6.23	5.49	4.03	1.10
Frustration	1.40	2.22	1.40	1.30
Overall	4.96	2.76	-	-

Fig. 10. NASA TLX questionnaire results.

weights (which adds up to 15). "Fig. 10" shows the measured individual ratings, their weights, and the overall rating. It is worth mentioning that the ratings are measured in the 0-20 range and the weights are in the 0-5 range.

B. Correlations

All gathered data were analyzed for correlations by examining their Pearson correlation coefficients. Only correlations significant at the 0.05 level (2-tailed) were considered. A statistically significant correlation was found between the Overall ratings of both methods ($r = .541$, $p = 0.004$). Our interpretation is that participants have various standards and they rated the lip animations relative to these standards. They also consistently rated our method higher relative to the baseline method.

All ratings of our method (Natural, Robotic, Artificial, Immersion-Breaking, Quality, Temporally Synchronized,Visually Matching) were pairwise correlated at least at the 0.005 level, showing that participants were consistent in their answers. Similarly, all of these ratings for the baseline method were mutually correlated at the 0.01 level, except for the pair Robotic and Temporally Synchronized. This exception makes sense, given that both methods used identical timestamps, and thus the temporal synchronization was the same. Users correctly identified that despite good temporal synchronization, the lips moved unnaturally fast with the baseline method—which could be perceived as robotic. Multiple users also pointed this out in their written feedback.

Correlations were also present between the two methods, especially in the Temporal Synchronization rating ($r = .768$, $p = 0.000001$) as is explained in the previous paragraph. The ratings of the last three aspects (Quality, Temporally Synchronized, Visually Matching) were all correlated across the two methods at the 0.01 level. Exceptions were the Immersion-Breaking rating and the Robotic rating, which were not correlated with any of the ratings of the other method. Participants who achieved a higher Presence rating based on their questionnaire answers were more likely to rate our method favorably, as evidenced by the strong correlation found between the Presence rating and the Overall rating of our method ($r = .571$, $p = 0.001$).

Presence correlated with each of the aspects that make up the Overall rating at the 0.05 significance level. Such correlation was not found between Presence and the Overall rating of the baseline method (r = .246, p = 0.191) or any of its sub-ratings. This could suggest that our method holds up better under closer inspection. The same could be said about participants with a high Emotion rating from the questionnaire, given the correlation found between the Emotion rating and the Overall rating of our method (r = .467, p = 0.009). Again, no such correlation was found with the rating of the baseline method (r = .256, p = 0.172).

Correlations were found between some of the aspects measured by the questionnaire – namely Presence, Immersion and Flow – showing that they are closely related. The strongest correlation was between Immersion and Flow (r = .580, p = 0.001), followed by Presence and Flow (r = .436, p = 0.016), and finally Presence and Immersion (r = .413, p = 0.023). Finally, no significant correlations were found between the daily computer use rating and other measured quantities.

C. Qualitative Comments

At the end of the evaluation, each participant was asked to give optional feedback, both on the lip-sync animation and on the experiment itself. While most participants praised the animations both verbally and in the feedback sheet, there have also been some valid points of criticism.

A number of participants suggested that the experiment should have included a practice stage, where they could familiarize themselves with the virtual environment and get a feel for the baseline lip-sync quality, arguing that they had nothing to compare the first phrase to. Another interesting point that was mentioned is that the lip motion seemed too isolated from the rest of the face, mostly because of the lack of emotion—they reported that some of the phrases sounded like they carry emotion, while the facial expression remained neutral.

After they correctly identified that they were comparing between only two lip-sync methods, some participants expressed that they wished the number of methods had been mentioned in advance. However, this information had been intentionally withheld until the end of the experiment, with the intention of minimizing bias. Finally, some participants reported that there was too much paperwork involved in the experiment, and several other participants have expressed mild displeasure with the amount of paperwork verbally. However, none of the participants were discouraged enough to withdraw from the experiment.

VI. Discussion

Our approach has multiple advantages over other approaches. First of all, co-articulation is not limited to neighboring visemes. Some solutions only allow co-articulation between pairs or triads of phonemes/visemes, even though in real speech, co-articulation can affect phonemes that are up to 5 units apart [32]. In our solution, important poses (such as bilabial stops) do not get broken due to co-articulation of neighboring visemes. This can be a problem with the approach based on dominance functions [4].

The output is in the form of animation curves. Animation curves can easily be edited by an animator, allowing for further refinement. Some other approaches control the facial rig at run-time, which also means that there are no animations to be manually adjusted by an animator. Moreover, our approach only requires an animator to set up 15 viseme poses with constraints. Some other techniques require an animator to prepare several hundred animations for transitioning between visemes [28].

Our algorithm can generate speech animations quickly enough to allow for real-time editing with a live preview. Some other approaches are based on iterative algorithms [29], which take longer to evaluate. The generated motion curves are at least C1 continuous. The motion produced by some other approaches (such as negative exponential dominance functions [4]) is only C0 continuous, and its discontinuities at the C1 level could be perceived as unnatural [27]. Furthermore, no training data set is needed. Our approach simulates co-articulation based on a set of pre-defined constraints. With data-driven approaches that learn explicit transitions between pairs—or longer sequences—of visemes from recorded data [5], the output quality tends to be limited by the size of the training data set.

In terms of limitations, our approach currently only supports neutral speech. It could be extended to analyze speech audio for tone of speech and adjust the animation accordingly. The jaw and lip parameters produced by the lip-sync solution [1] could be applied onto jaw-related and lip related parameters in our model. There is a lot more to facial animation than just lip motion. Moreover, a simulation of gaze could be used to add credibility to the lip-sync. To be truly convincing, the lip-sync solution should also be paired with a simulation of head motion that adds emphasis where necessary. Facial expressions, including eyebrow and cheek motion, can add emotional depth to the speech animations. Body language and gestures play an important role as well.

VII. Conclusion

In this paper, we have explored an alternative solution for simulating co-articulation. Our constraint-based model has accomplished the goal of producing more natural lip-sync animations than standard keyframe interpolation techniques. In the evaluation, our model significantly outperformed the baseline method. While the quality of our output animations is already satisfactory (it was rated as 70% natural by the participants in the user study), it is still not on par with the performance capture or professional hand-made animation. However, our method produces animation curves that are easily editable by an animator, which allows for further refinement. Using animations generated by our approach as a starting point saves a significant amount of time compared to creating lip-sync animations manually from scratch, and can lead to high-quality results more quickly.

ACKNOWLEDGMENTS

This research was partially supported by the project that has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 739578 and the Government of the Republic of Cyprus through the Deputy Ministry of Research, Innovation and Digital Policy. Authors would like to thank all participants that took part in the user study.

REFERENCES

[1] P. Edwards, C. Landreth, E. Fiume, and K. Singh, "Jali: An animator-centric viseme model for expressive lip synchronization," *ACM Trans. Graph.*, vol. 35, no. 4, Jul. 2016. [Online]. Available: https://doi.org/10.1145/2897824.2925984

[2] T. Karras, T. Aila, S. Laine, A. Herva, and J. Lehtinen, "Audio-driven facial animation by joint end-to-end learning of pose and emotion," *ACM Trans. Graph.*, vol. 36, no. 4, Jul. 2017. [Online]. Available: https://doi.org/10.1145/3072959.3073658

[3] F. I. Parke and K. Waters, *Computer Facial Animation*, 2nd ed. AK Peters Ltd, 2008.

[4] M. Cohen and D. Massaro, "Modeling coarticulation in synthetic visual speech," 03 1999.

[5] S. L. Taylor, M. Mahler, B.-J. Theobald, and I. Matthews, "Dynamic units of visual speech," in *Proceedings of the ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, ser. SCA '12. Goslar, DEU: Eurographics Association, 2012, p. 275–284.

[6] L. Williams, "Performance-driven facial animation," in *ACM SIGGRAPH 2006 Courses*, ser. SIGGRAPH '06. New York, NY, USA: Association for Computing Machinery, 2006, p. 16–es. [Online]. Available: https://doi.org/10.1145/1185657.1185856

[7] D. Hendler, L. Moser, R. Battulwar, D. Corral, P. Cramer, R. Miller, R. Cloudsdale, and D. Roble, "Avengers: Capturing thanos's complex face," in *ACM SIGGRAPH 2018 Talks*, ser. SIGGRAPH '18. New York, NY, USA: Association for Computing Machinery, 2018. [Online]. Available: https://doi.org/10.1145/3214745.3214766

[8] Y. Zhou, Z. Xu, C. Landreth, E. Kalogerakis, S. Maji, and K. Singh, "Visemenet: Audio-driven animator-centric speech animation," *ACM Trans. Graph.*, vol. 37, no. 4, Jul. 2018. [Online]. Available: https://doi.org/10.1145/3197517.3201292

[9] C. Cao, D. Bradley, K. Zhou, and T. Beeler, "Real-time high-fidelity facial performance capture," *ACM Trans. Graph.*, vol. 34, no. 4, Jul. 2015. [Online]. Available: https://doi.org/10.1145/2766943

[10] L. Ma and Z. Deng, "Real-time hierarchical facial performance capture," in *Proceedings of the ACM SIGGRAPH Symposium on Interactive 3D Graphics and Games*, ser. I3D '19. New York, NY, USA: Association for Computing Machinery, 2019. [Online]. Available: https://doi.org/10.1145/3306131.3317016

[11] S. Laine, T. Karras, T. Aila, A. Herva, S. Saito, R. Yu, H. Li, and J. Lehtinen, "Production-level facial performance capture using deep convolutional neural networks," in *Proceedings of the ACM SIGGRAPH / Eurographics Symposium on Computer Animation, Los Angeles, CA, USA, July 28-30, 2017*, J. Teran, C. Zheng, S. N. Spencer, B. Thomaszewski, and K. Yin, Eds. Eurographics Association / ACM, 2017, pp. 10:1–10:10. [Online]. Available: https://doi.org/10.1145/3099564.3099581

[12] C. Bregler, M. Covell, and M. Slaney, "Video rewrite: Driving visual speech with audio," in *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques*, ser. SIGGRAPH '97. USA: ACM Press/Addison-Wesley Publishing Co., 1997, p. 353–360. [Online]. Available: https://doi.org/10.1145/258734.258880

[13] M. Brand, "Voice puppetry," in *Proceedings of the 26th Annual Conference on Computer Graphics and Interactive Techniques*, ser. SIGGRAPH '99. USA: ACM Press/Addison-Wesley Publishing Co., 1999, p. 21–28. [Online]. Available: https://doi.org/10.1145/311535.311537

[14] S. Kshirsagar and N. Magnenat-Thalmann, "Visyllable based speech animation," *Computer Graphics Forum*, vol. 22, no. 3, pp. 631–639, 2003. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/1467-8659.t01-2-00711

[15] Y. Cao, P. Faloutsos, E. Kohler, and F. Pighin, "Real-time speech motion synthesis from recorded motions," in *Proceedings of the 2004 ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, ser. SCA '04. Goslar, DEU: Eurographics Association, 2004, p. 345–353. [Online]. Available: https://doi.org/10.1145/1028523.1028570

[16] Z. Deng, U. Neumann, J. P. Lewis, T. Kim, M. Bulut, and S. S. Narayanan, "Expressive facial animation synthesis by learning speech coarticulation and expression spaces," *IEEE Trans. Vis. Comput. Graph.*, vol. 12, no. 6, pp. 1523–1534, 2006. [Online]. Available: https://doi.org/10.1109/TVCG.2006.90

[17] S. Deena, S. Hou, and A. Galata, "Visual speech synthesis by modelling coarticulation dynamics using a non-parametric switching state-space model," in *International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction*, ser. ICMI-MLMI '10. New York, NY, USA: Association for Computing Machinery, 2010. [Online]. Available: https://doi.org/10.1145/1891903.1891942

[18] S. Deena, S. Hou, and A. Galata, "Visual speech synthesis using a variable-order switching shared gaussian process dynamical model," *IEEE Transactions on Multimedia*, vol. 15, no. 8, pp. 1755–1768, 2013.

[19] S. Asadiabadi, R. Sadiq, and E. Erzin, "Multimodal speech driven facial shape animation using deep neural networks," in *2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2018, pp. 1508–1512.

[20] S. Taylor, T. Kim, Y. Yue, M. Mahler, J. Krahe, A. G. Rodriguez, J. Hodgins, and I. Matthews, "A deep learning approach for generalized speech animation," *ACM Trans. Graph.*, vol. 36, no. 4, Jul. 2017. [Online]. Available: https://doi.org/10.1145/3072959.3073699

[21] A. Thangthai, B. Milner, and S. Taylor, "Synthesising visual speech using dynamic visemes and deep learning architectures," *Computer Speech & Language*, vol. 55, pp. 101–119, 2019. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0885230818300275

[22] F. I. Parke, "Parameterized models for facial animation," *IEEE Comput. Graph. Appl.*, vol. 2, no. 9, p. 61–68, Sep. 1982. [Online]. Available: https://doi.org/10.1109/MCG.1982.1674492

[23] B. Uz, U. Gudukbay, and B. Ozguc, "Realistic speech animation of synthetic faces," in *Proceedings Computer Animation '98 (Cat. No.98EX169)*, 1998, pp. 111–118.

[24] P. Aleksic, G. Potamianos, and A. Katsaggelos, "10.8 – exploiting visual information in automatic speech processing," 2005.

[25] A. Löfqvist, "Speech as audible gestures," in *Speech Production and Speech Modelling. NATO ASI Series (Series D: Behavioural and Social Sciences)*, vol. 55. Springer, 1990.

[26] P. Cosi, E. M. Caldognetto, G. Perin, and C. Zmarich, "Labial coarticulation modeling for realistic facial animation," in *Proceedings. Fourth IEEE International Conference on Multimodal Interfaces*, 2002, pp. 505–510.

[27] S. A. King and R. E. Parent, "Creating speech-synchronized animation," vol. 11, no. 3, p. 341–352, May 2005. [Online]. Available: https://doi.org/10.1109/TVCG.2005.43

[28] Y. Xu, A. W. Feng, S. Marsella, and A. Shapiro, "A practical and configurable lip sync method for games," in *Proceedings of Motion on Games*, ser. MIG '13. New York, NY, USA: Association for Computing Machinery, 2013, p. 131–140. [Online]. Available: https://doi.org/10.1145/2522628.2522904

[29] O. M. Lazalde, S. Maddock, and M. Meredith, "A constraint-based approach to visual speech for a mexican-spanish talking head," *Int. J. Comput. Games Technol.*, vol. 2008, Jan. 2008. [Online]. Available: https://doi.org/10.1155/2008/412056

[30] K. Stevens, *Acoustic Phonetics*, 01 2000, vol. 109, pp. 607–607.

[31] J. Osipa, *Stop Staring: Facial Modeling and Animation Done Right*, 3rd ed. USA: SYBEX Inc., 2010.

[32] R. Kent and F. Minifie, "Coarticulation in recent speech production models," *Journal of Phonetics*, vol. 5, no. 2, pp. 115–133, 1977. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0095447019311234

[33] K. Tcha-Tokey, O. Christmann, E. Loup-Escande, and S. Richir, "Proposition and Validation of a Questionnaire to Measure the User Experience in Immersive Virtual Environments," p. 28, 2016.

[34] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research," in *Advances in Psychology*, ser. Human Mental Workload, P. A. Hancock and N. Meshkati, Eds. North-

Holland, Jan. 1988, vol. 52, pp. 139–183. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0166411508623869