Text-driven Speech Animation with Emotion Control

Wonseok Chae[1] and Yejin Kim[2*]
[1] Content Validation Research Section, Electronics and Telecommunications Research Institute
Daejeon, 34129 - Republic of Korea
[e-mail: wschae@etri.re.kr]
[2] School of Games, Hongik University
Sejong, 30015 - Republic of Korea
[e-mail: yejkim@hongik.ac.kr]
*Corresponding author: Yejin Kim

Received March 2, 2020; revised June 22, 2020; accepted July 18, 2020; published August 31, 2020

Abstract

In this paper, we present a new approach to creating speech animation with emotional expressions using a small set of example models. To generate realistic facial animation, two example models called key visemes and expressions are used for lip-synchronization and facial expressions, respectively. The key visemes represent lip shapes of phonemes such as vowels and consonants while the key expressions represent basic emotions of a face. Our approach utilizes a text-to-speech (TTS) system to create a phonetic transcript for the speech animation. Based on a phonetic transcript, a sequence of speech animation is synthesized by interpolating the corresponding sequence of key visemes. Using an input parameter vector, the key expressions are blended by a method of scattered data interpolation. During the synthesizing process, an importance-based scheme is introduced to combine both lip-synchronization and facial expressions into one animation sequence in real time (over 120Hz). The proposed approach can be applied to diverse types of digital content and applications that use facial animation with high accuracy (over 90%) in speech recognition.

Keywords: Speech animation, lip-synchronization, emotional expressions, facial expression synthesis, example models

A preliminary version of this paper was presented at ICONI 2019 and was selected as an outstanding paper.
This version includes detailed descriptions on the proposed approach and additional experimental results.
This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No. 2020-0-00437, Proactive interaction based VRAR virtual human object creation technology) and by Basic Science Research Program through the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2017R1C1B5017000).

http://doi.org/10.3837/tiis.2020.08.018 **ISSN : 1976-7277**

1. Introduction

Virtual characters can play an important role in the interaction between humans and computers. The human-computer interaction approach using human voices and visual-based facial animation not only generates users' interest but also allows users to concentrate on tasks. The human face is an important part of the body that expresses emotions as well as identification during conversations. Through the face of a virtual character who speaks naturally with rich expression, a user can feel as if the computer is a person from an interaction with a computer. For example, a lively speaking virtual character can provide information through an agent system. In recent years, 5G wireless communication has attracted much attention in social network services using avatars, which are virtual characters expressing lip-synced animations with emotional expressions in a virtual reality environment.

Creating animated facial expressions that speak naturally like real people is a difficult task. It should be able to express the shape of the lips that is precisely synchronized with the speaking voice. Numerous studies have presented ways to create visual speech animation with the speech track [1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11] while other approaches have focused on simulating facial movements from a set of physical properties [12, 13, 14, 15] or synthesizing emotional expressions from given facial models [16, 17, 18, 19, 20]. For more natural and realistic facial animation, an explicit solution is needed to combine the lip movements and facial expressions into one animation sequence.

In this paper, we present a new approach to generate convincing animated facial expression by combining lip-synced animation and emotional expressions of 3D face models. In our approach, three important steps are addressed: First, a simple and effective method is introduced to generate lip-synced animation with coarticulation effects from input text. Next, a blending approach based on scattered data interpolation is presented to synthesize facial expressions from given example models. Last, an importance-based technique is proposed to compose a final animation from a sequennce of lip-synced and emotional expression models in real time (over 120Hz). As shown in our experimental results, the resulting animations are smooth and expressive with high accuracy (over 90%) in speech recognition while they are intuitive to produce using the provided interface.

The remainder of this paper is divided as follows. Previous approaches for speech and emotional animation generation are reviewed in Section 2. The main approach for text-based speech generation using emotion control is described in Section 3. The experimental results for generating various speech animations in real time are demonstrated in Section 4. We conclude this paper with potential improvements in Section 5.

2. Related Work

2.1 Text-driven Speech Animation Generation

There have been active research works proposed for generating text-driven speech animation. In these methods, facial animations are synchronized with a speech source and produced from 2D or 3D sample data. For exmaple, the 2D methods focused on lip synchronization by analyzaing input videos or image samples [1, 2, 3, 4]. Using a statistical training model, Brand presented speech animation generation which added upper-face expressions to a speech

animation driven by an audio track [5]. However, this type of approach is not easy for general users to control expressions and to create desired animation results.

On the other hand, speech animation generation using 3D models provided users more flexiblity and realism in animation production. Parke introduced an early parametric model for facial expressions [6]. Based on this model, Pearce et al. presented a 3D speech animation system that converted input phonemes into time-varying control parameters and produced an animation sequence. Kalberer et al. extended this approach to produce more realistic speech animation by capturing a speeching human face from a 3D digitizer [8]. In more recent works, Wang et al. utilized the data obtained from electromagnetic articulography (EMA) devices which measure the tongue and mouth movements during human speech [9]. In their approach, a 3D model followed speaking words at the phoneme label. However, they mainly focused on accurate pronunciation for learning language and did not produce animations from the viseme label. To create more realistic speech animation, Kawai et al. examined out the movement of the teeth and tongue along with the movement of the lips [10]. They focused on the tip of the tongue between the teeth or the back of the tongue, but this type of approach requires sufficient data collected from the subject in advance to produce animation results. Kuhnke and Ostermann collected a sequence of 3D mesh data along the phoneme label by capturing 3D facial movement and recording voice data at the same time [11]. With a regression-based method, they presented a combination method of several speech features for better performance. However, it required a sequence of high-speed images from a live human performance, which was time-consuming and costly. Most of these parameter-based approaches with 3D models have only focused on lip synchronization. Futhermore, they are not suitable for general users to create desired facial expressions by directly controlling a set of parameters.

2.2 Emotional Expression Animation

Research on creating whole-face expressions without a speech source have been also actively studied over decades. Physically-based methods synthesized facial expressions by simulating the physical relationship between facial skin and muscles. Lee et al. introduced a biomechanical model for facial animation based on the range scanned data [12]. Sifakis et al. simulated a physical model by solving the inverse activation problem [13]. Cong et al. proposed an anatomical face model that can adopt to different character models [14]. Ichim et al. presented a physics-based approach by simulating muscle, skin, and bone interaction underneath a face [15]. However, most physics-based approaches require a muscle activation model with an optimization solver, which is not suitable for real-time animation generation.

In data-driven methods, a set of facial motion data captured from a live actor are used to produce realistic animation. Using 2D photographs, Pighin et al. introduced a 3D shape morphing technique to show photorealistic facial expressions [16]. Guenter et al. introduced a performance-driven system for capturing both 3D geometry and color data using a large set of marker points attached to the face [17]. Ju and Lee showed a similar system using Markov random fields that were simulated at two levels [18]. Using a maximum a posteriorir (MAP) framework, Lau et al. presented an interaction system that created various expression models via a sketch device [19]. Barrielle et al. proposed a blending system that deformed a face model based on facial dynamics [20]. However, most of these approaches have mainly focused on emotional expression synthesis and do not address an automated way of lip-synchronization with a speech source.

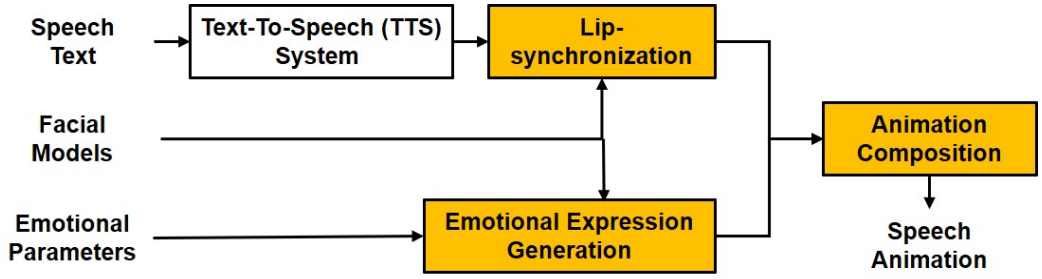

Fig. 1. System overview.

2.3 Emotional Speech Animation

To create facial animation with speech and emotion, Jia et al. mapped various human emotions to a 3D parameter space and generated audio-visual speech animation by converting neutral speech to emotional speech [21]. In this approach, some key emotions, such as relaxation and fear [22], were not well expressed due to the ambiguity in the emotional speech conversion using a Boosting-Gaussian mixture model (GMM). Wan et al. presented a cluster adaptive training (CAT) method which made it possible to independently control the mouth shape and facial expressions [23]. Although their approach could generate an emotional speech animation through a simple interface of emotion control, training a CAT model requires the collection of a speech and video corpus, which is laborious, whenever new emotion is introduced. Stef et al. introduced an apporoach that converted a given text into the international phonetic alphabet (IPA), mapped the corresponding lip shape to each symbol, and generated emotional animation by a key-framing technique [24]. Their approach relied on a commerical animation tool and it was necessary to adjust a large number of emotional parameters to create a desired expression.

3. Approach

3.1 Overview

Fig. 1 shows an overview of the proposed approach which consists of three generative steps: lip-synchronization, emotional expression generation, and animation composition. First, a speech animation is synchronized with an input speech track which is synthesized from an existing text-to-speech (TTS) system. With the given input text, the TTS system produces a speech track with corresponding phonemes and their lengths. A set of visual counterparts of phonemes called *visemes* [25] are prepared with 3D face models as example models. Similarly, a set of 3D emotional models are used to represent basic facial expressions. Using a sequence of key viseme models as key frames, our approach generates a speech animation by interpolating the key frames and synchronizing the animation with the phoneme sequence. During this synthesis, coarticulation effects are considered by adjusting the viseme model at each key frame based on the corresponding phoneme length.

Given emotion control of key expressions on a parameter space, our approach produces an emotional expression at the same time with the speech animation generation. At each frame, a new expression model is interpolated from a set of key emotional models and their parameters obtained from the emotion space, which parameterizes the key expressions on a 2D space. For

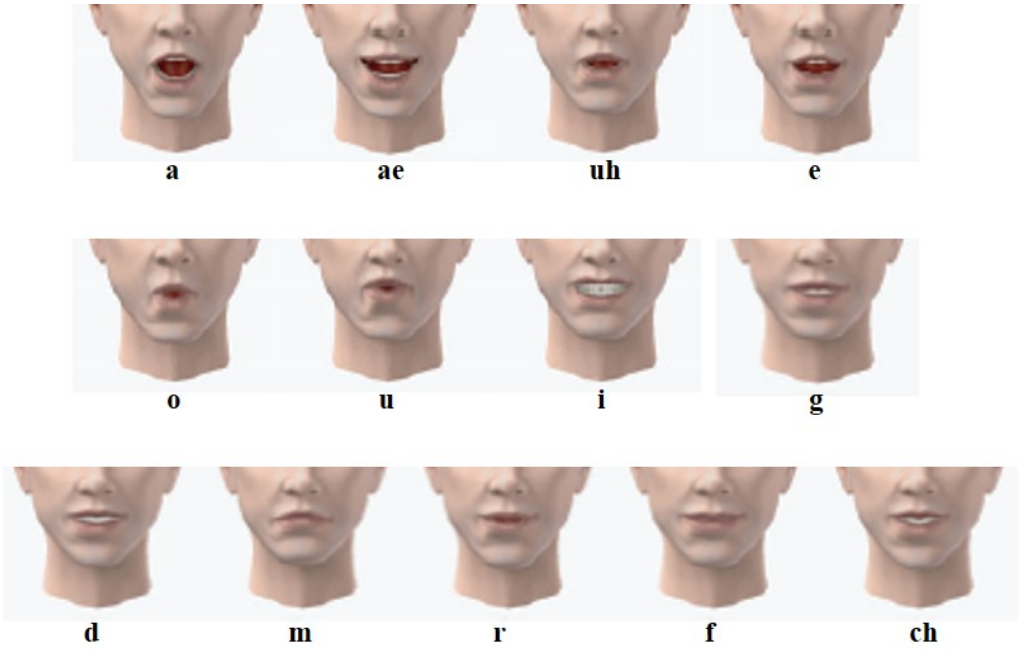

Fig. 2. Key viseme models used for the consonants and the vowels.

real-time peformance, our approach adopts a scattered data interpolation techinque to synthesize the emotional expression model from the given parameter vector obtained from the emotion space.

The final animation is generated by combing the lip-synced and expression sequences at each frame. For a natural looking animation, an importance-based approach is introduced to combine the two models, viseme and expression, into one output animation without conflicts. In this method, an importance value is assigned to each vertex of the face model based on its relative contribution to the accurate pronunciations with respect to the emotional expressions. Using the importance values as weights, each vertex of the composite model is decided by blending the corresponding vertex of the viseme model and the expression model in real time.

3.2 Lip-synchronization

This section describes our method for generating a speech animation from key viseme models and synchronizing the animation with an input speech track. To obtain the speech track with the sequence of phonemes and their lengths, an existing TTS system was utilized to synthesize the speech track because the recording of a human voice requires a labeling process which is laborious and time-consuming for users. Given the input speech track with phoneme information, we generate the speech animation from the key viseme models and synchronize the animation with the phoneme sequence.

To generate natural looking lip movements, a set of key viseme models was prepared for unique sounds (aka phonemes). In American English, there are approximately 44 phonemes, which can be represented by the smaller number of viseme models because one viseme can represent more than one phoneme. We assumed that those models could be created by blending a small set of key visemes. After careful observation of visual phonemes, we selected 14 key visemes for our approach: six of them represent the consonantal phonemes, seven of

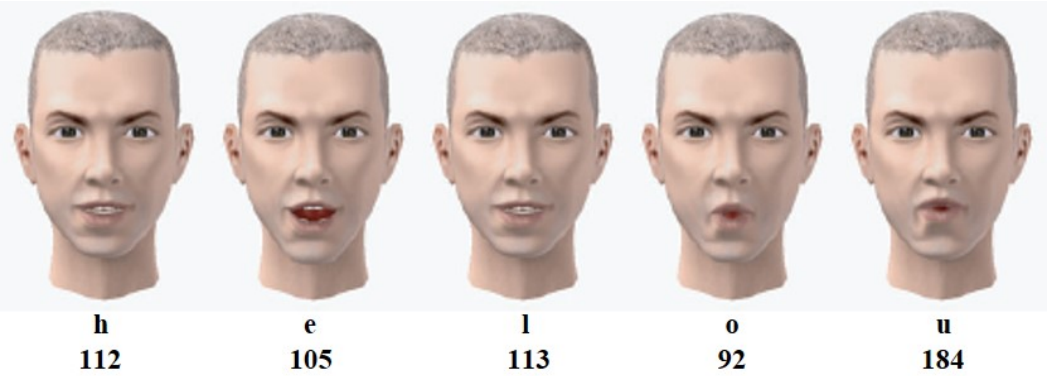

Fig. 3. Key viseme models with their durations for a sequence, "Hello".

them represent the vocalic phonemes, and one added viseme represents neutrality or silence, as shown in **Fig. 2**. Given a neutral face model, all the viseme models can be produced by deforming the shape of lips on the neutral model.

Given the viseme set and the input phoneme sequence, viseme models can be selected for corresponding phonemes as shown in **Fig. 3**. Let $\boldsymbol{S}$ and $\boldsymbol{L}$ be the phoneme sequence and the phoneme length, respectively, where $\boldsymbol{S} = \{S_1, S_2, \dots, S_m\}$ and $\boldsymbol{L} = \{l_1, l_2, \dots, l_m\}$. Let $\boldsymbol{P}$ be the corresponding viseme sequence, where $\boldsymbol{P} = \{P_1, P_2, \dots, P_m\}$. Here, P_j is a polygonal mesh and composed of a set of vertices $\mathrm{v}_i^j = (x_i^j, y_i^j, z_i^j)$, where $1 \le i \le n$. Given $\boldsymbol{L}$, we first locate each key viseme P_j at the time instance t_j, where $1 \le j \le m$, along the time axis, $\boldsymbol{T}$, where $\boldsymbol{T} = \{t_1, t_2, \dots, t_m\}$. If P_j is placed at a local extremum, which is achieved at every key frame, t_j for P_j can be estimated as follows:

$$t_j = \frac{l_j}{2} + \sum_{k=1}^{j-1} l_k. \tag{1}$$

Here, we assumed that each P_j is placed at the middle of the corresponding l_j along $\boldsymbol{T}$.

With P_j placed at the corresponding time t_j, we first use the Catmull-Rom spline interpolation method to construct piece-wise cubic splines that interpolate each sequence of the corresponding v_i^j for the selected P_j. To ensure the local extremum at the key frame, the tangent vector of the cubic splines is set to be zero at every key frame. The x coordinates for v_i at t, where $t_j \le t \le t_{j+1}$, is represented by the cubic polynomial of the curve, $x_i(t) = at^3 + at^3 + ct + d$. Here, the coefficients (a, b, c, d) are estimated from the following constraints: $x_i(t_j) = x_i^j, x_i(t_{j+1}) = x_i^{j+1}$, and $x_i'(t_j) = x_i'(t_{j+1}) = 0$, where x_i^j and x_i^{j+1} are the x coordinates of v_i at times t_j and t_{j+1}, respectively. The y and z coordinates can be estimated in the same manner.

Next, for S_j with a large l_j, where the lip movement needs to be maintained for a while, we assign a viseme interval M_j centered at t_j and define its length $l(M_j)$ as follows:

$$l(M_j) = \begin{cases} l_j - \delta_1, & l_j > \delta_1 \\ 0, & \text{otherwise} \end{cases}. \tag{2}$$

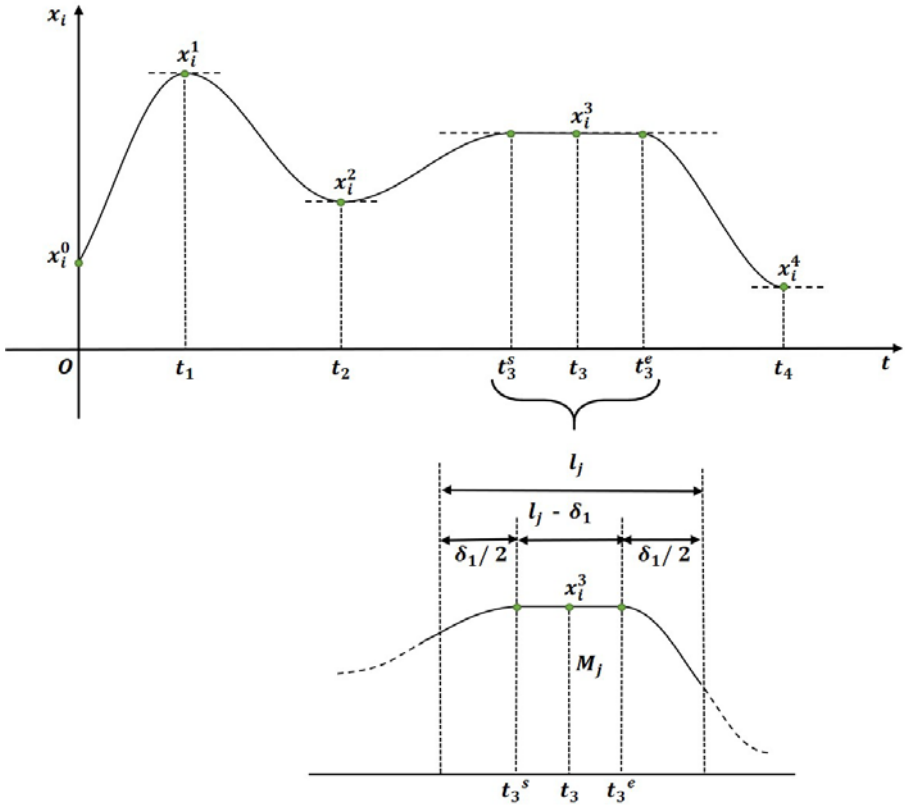

Fig. 4. A piece-wise cubic spline interpolation with a viseme maintenance interval for the *x* coordinate.

Here, if $l_j > \delta_1$, both ends of M_j, t_j^s and t_j^e, are used instead of t_j for spline interpolation with the nearby key visemes. **Fig. 4** shows an example with phoneme S_3 at x_3.

Finally, the coarticulation effects are considered during the lip-synced animation generation. In general, when a person makes a speech, one does not always show a complete sequence of the visemes as some phoneme lengths are very short. This forces the lip shape of the current viseme to be like the neighboring phonemes, which is called *coarticulation*. To visualize this effect, the current viseme model is constrained by its phoneme length and the shape of the successive visemes. Starting from P_1 in $\boldsymbol{P}$, each P_j is adjusted successively with respect to the previous viseme model, to generate a new viseme sequence $\bar{\boldsymbol{P}}$, where $\bar{\boldsymbol{P}} = \{\bar{P}_1, \bar{P}_2, \ldots, \bar{P}_m\}$. Here a vertex position $\bar{v}_i^j$ for $\bar{P}_j$ can be estimated as follows:

$$\bar{v}_i^j = w(l_j) \times v_i^j + \left(1 - w(l_j)\right) \times v_i^{j-1}, \tag{3}$$

where $w(l_j)$ is a weight function of the corresponding l_j. In our approach, we set $w(l_j) = 0$, if $l_j > 120$ms, which is empirically set, while S_j is generated between 30ms and 400ms from the TTS system. To achieve a smooth transition between two visemes in the animation sequence, the transition speed increases gradually at start, abruptly at middle, and slows down at end as

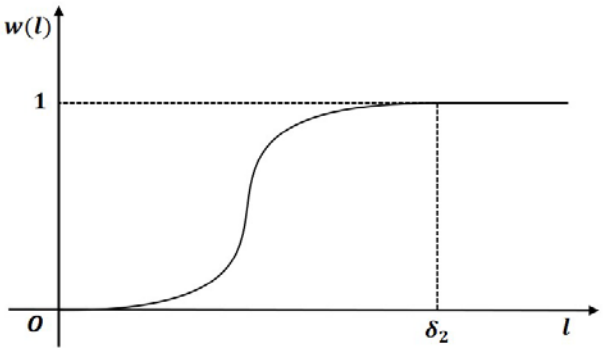

Fig. 5. A weight function for coarticulation effect.

shown in **Fig. 5**. Based on this empirical observation, the weight function $w(l_j)$ is defined as follows:

$$w(l_j) = \begin{cases} -\frac{2l_j^3}{\delta_2^3} + \frac{3l_j^2}{\delta_2^2}, & l_j > \delta_1, \\ 1, & \text{otherwise} \end{cases} \tag{4}$$

which is derived from the constraints: $w(0) = 0$, $w(\delta_2) = 1$, and $w'(0) = w'(\delta_2) = 0$. This way, the modified viseme sequence $\bar{\boldsymbol{P}}$ exhibits a smooth and natural looking lip-synchronization.

3.3 Emotional Expression Generation

In this section, we explain an emotional expression synthesis from a set of example models and emotion control. Like viseme models for lip-synced animation generation, various emotional expressions can be constrcuted by blending a set of key expression models. Using the emotion space diagram [26], our approach adopted the following emotional expressions: happy, sad, angry, afraid, surprised, and neutral expressions. Their 3D face models and parameter space for the expression synthesis are shown in **Fig. 6** and **Fig. 7**, respectively. Like the viseme set, the neutral expression model is the base model, representing silence.

Given the emotion control on the parameter space, our approach adopted a multidimensional scattered data interpolation technique [27, 28] due to its real time performance and suitability for a small set of example models. In this method, a weight function $w_i(\cdot)$ for the *i*th key expression is predefined with a combination of linear and radial basis functions, $A(\cdot)$ and $R(\cdot)$. Here, $A(\cdot)$ approximaes the global shape of weight functions while $R(\cdot)$ adjusts the global shape locally to exactly interpolate the key expression models. Let **p** be a parameter vector derived from the continuous curve defined on the parameter space as shown in **Fig. 7**. At runtime, while **p** is specified interactively by a user, the corresponding expression model is generated by blending the key expressions with respect to the weight values obtained from $w_i(\cdot)$ at **p**.

Let M be the number of key expression models located in the parameter space. For each expression model E_i, where $1 \leq i \leq M$, $w_i(\boldsymbol{p})$ is estimated as follows:

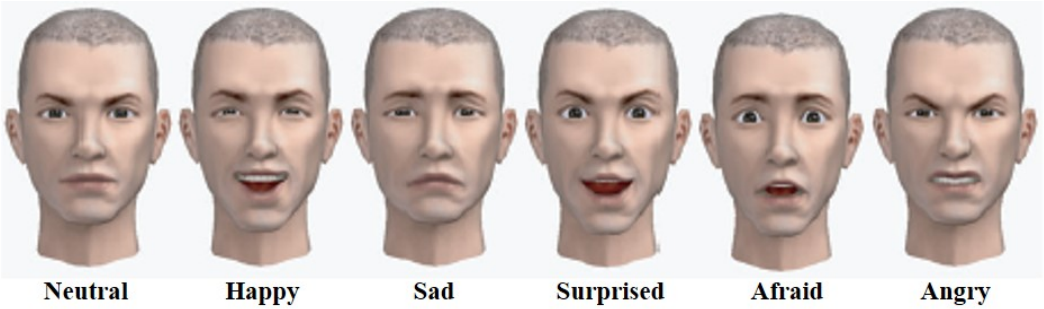

Fig. 6. Key expression models for emotion.

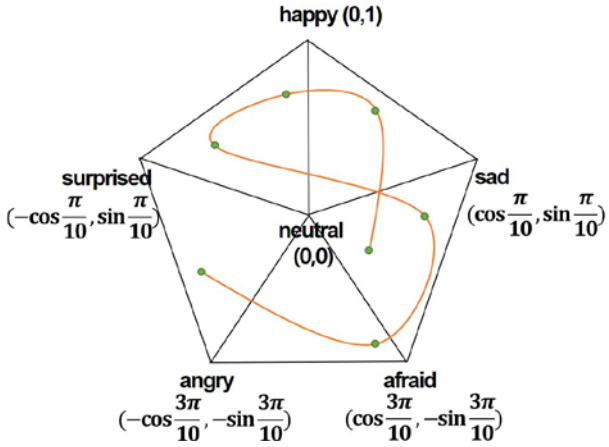

Fig. 7. The parameter space for expression synthesis: A cubic spline curve is used to continuously feed the emotional parameters during animation synthesis.

$$w_i(\boldsymbol{p}) = \sum_{l=0}^{2} a_{il} A_l(\boldsymbol{p}) + \sum_{j=1}^{M} r_{ji} R_j(\boldsymbol{p}), \tag{5}$$

where a_{il} and r_{ji} are the coefficients of linear basis functions and radial basis functions, respectively. Let $\boldsymbol{p}_i$, where $1 \leq i \leq M$, be the parameter vector of E_i. To interpolate the key expression model exactly, the weight of E_i is $w_i(\boldsymbol{p}_j) = 1$ for $i = j$ and $w_i(\boldsymbol{p}_j) = 0$ for $i \neq j$. The unknown coefficients a_{il} of the linear bases and the unknown coefficients r_{ji} of the radial bases can be estimated by employing a least squares method and computing the residuals for the key expressions, respectively [28].

With the weight functions predefined, the key expression models can be blended at runtime. Using w_j for $E_j, 1 \leq j \leq M$, a new expression model E_{new} at an input parameter vector $\boldsymbol{p}_{in}$ is generated as follows:

$$v_i^{new}(\boldsymbol{p}_{in}) = \sum_{j=1}^{M} w_j(\boldsymbol{p}_{in}) v_i^j, \tag{6}$$

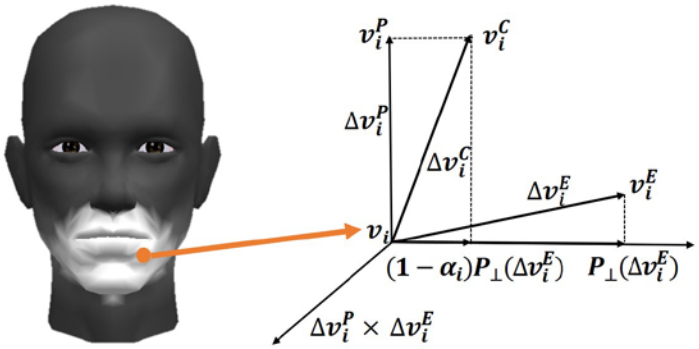

Fig. 8. Composition of vertex displacements based on the expression and viseme importance: The brighter regions indicate higher importance values [28].

Fig. 9. Composition of viseme model (consonant 'm') and expression model (happy).

where v_i^{new} and v_i^j, where $1 \leq i \leq n$, are the vertex of E_{new} and that of E_j, respectively. In practice, $\boldsymbol{p}_{in}$ is derived from a cubic spline curve in the parameter space to continuously feed the emtional paramters during animation synthesis. The curve shown in **Fig. 7** shows this example.

3.4 Animation Composition

In this section, we explain a combining process for a lip-synced animation and emotional expressions. As described in the previous sections, our approach generates a lip-synced animation and emotional expressions at the same time. That is, we need to compose a new 3D face model by combining a viseme model and an expression model for each frame. Based on careful observations, when there are conflicts between those two facial models, the vertex motion is mainly constrained by the lip movements of the viseme models. To combine them smoothly, we adopted an importance-based method to displace each face vertex using the pronunciation and emotional importance [28].

To combine a viseme model P and an expression model E, we define the displacement vectors $\Delta v_i^E = v_i^E - v_i$ and $\Delta v_i^P = v_i^P - v_i$ for every vertex v_i. Let $C = \{v_1^C, v_2^C, \ldots, v_n^C\}$ be the composite face model and $\Delta v_i^C = v_i^C - v_i$ be the displacement vector of a vertex $\Delta v_i^C \in C$. The v_i^C should lie on the plance spanned by Δv_i^E and Δv_i^P and containing v_i as shown in **Fig. 8**. For example, **Fig. 9** shows a composite model combined from a viseme model (consonant 'm') and a expression model (happy). It is noteworhty that the shape of the consonant model is

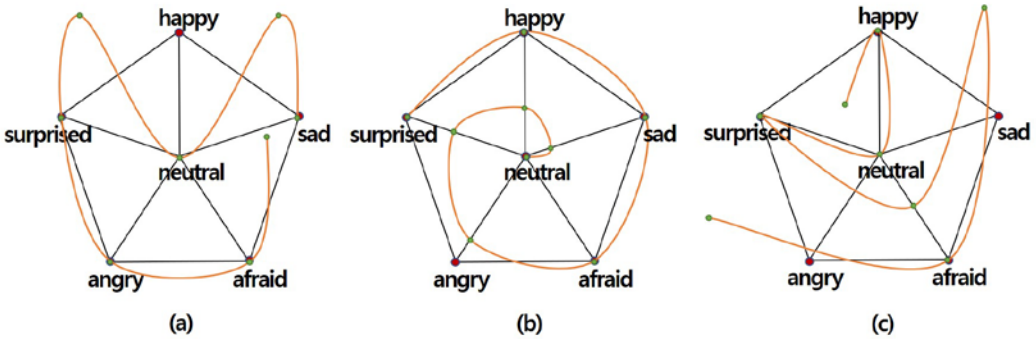

Fig. 10. Various emotional parameter curves.

Table 1. Models used for the experiments

Geometry			
Vertices	1192	1220	4522
Polygons	2194	2246	8982

Table 2. Input sentences

Quotes		Total Words
A	Gettysburg Address (Abraham Lincoln)	272
B	Inaugural Address (John F. Kennedy)	1,366
C	I Have a Dream (Martin Luther King Jr.)	1,667

preserved around the lips while the happy expression is preserved in other parts of the face model.

4. Experimental Results

Our experiments were conducted on an Intel i3 3.4GHz PC with 8GB memory and Nvidia GTX 1060 GPU with 3GB memory. A commercial TTS system called VoiceText [29] was used to produce a phoneme sequence from input text. Using a set of 3D face models for key visemes and expressions, our approach generates a emotional speech animation which is synchronized with a given speech track. As shown in **Fig. 10**, when a user controls the emotional parameter, our approach first generates a viseme model and an emotional expression model for each frame and then combines them into the final animation sequence in real time. Our experimental results are best understood by the online video located at https://drive.google.com/open?id=1JqUMFBjoN5ewhnpofIGduPs71XjACMKR.

Table 1 shows the three models (man, woman, and gorilla) used in our experiments with the polygonal information for each model. As shown in **Fig. 10**, various emotional parameter curves are used to generate the final animation sequences shown in **Fig. 11**. In **Fig. 11**, each row shows 12 sample frames of a lip-synced result, an emotional expression result, and their composite result generated for the three models. It is noteworthy that only the composite animation in the third row is displayed at runtime. Each composite model is generated by

Table 3. Time analysis using the text A (Gettysburg Address)

Computation Time	Man	Woman	Gorilla
Lip-synchronization (ms/frame)	0.0362	0.0378	0.1182
Emotional Expression Generation (ms/frame)	0.0212	0.0263	0.0742
Animation Composition (ms/frame)	0.0392	0.0402	0.1081
Total (ms/frame)	0.0966	0.1043	0.3005
Frame Rate (Hz)	132.2	130.8	128.8

Table 4. User case study: Speech recognition

Quotes	Total Words	Accuracy (%)		
		Speech Only	Visual Only	Visual Speech
A	96	82	21	93
B	92	84	22	94
C	102	86	19	91

combining the corresponding viseme model and the expression model shown in the same column. For the man model, the lip-synced animation was produced by using the input text A in **Table 2**. During the animation generation, the emotion expressions are concurrently synthesized by using the input parameter curve shown in **Fig. 10**(a). It is noteworthy that our expression synthesis technique can generate the expression models at parameter vectors even outside the parameter space spanned by the key expression models due to the linear functions used for weight function estimation. As shown in the final animation sequence, each composite model exhibits the key shape of lips and the emotional expression without conflicts. In the same manner, the text B in **Table 2** with the parameter curve in **Fig. 10**(b) and the text C with the parameter curve in **Fig. 10**(c) were used for the woman and gorilla models, respectively.

For efficiency analysis, each model is tested with the same input text sequence, which is a part of the text A in **Table 2**. The TTS system produced the corresponding phoneme sequence of 87 seconds long from the given text, which is composed of 272 words. The same emotion parameter curve in **Fig. 10**(a) was applied to all three models. **Table 3** shows performance statistics obtained from this test. The table shows the computation time (in ms per frame) for lip-synchronization, emotional expression generation, and animation composition without the rendering time. In the table, the frame rate indicates the number of frames per second to produce the final animation sequence with the rendering time. This test demonstrates that our approach is capable of a real-time performance (over 120Hz) to produce emotional speech animations with interative control of emotion from general users.

Finally, **Table 4** shows the effectiveness of our approach in speech recognition. In this test, 10 native English speakers are hired and tested with three input texts, which are parts of the quotes in **Table 2**. Each participant takes dictation from speech sound (without visual animation), visual animation (without speech sound), and visual speech animation, respectively. The accuracy of dictation for each input text is averaged from all participants and compared in **Table 4**. It demonstrates that the visual speech animation shows over 90% accuracy in speech recognition, effective in communication.

In sum, our approach can generate a speech animation with emotions, which does not rely on large datasets (i.e. hundredths of speech sentences and tenths of emotions captured from people) required for a statistical model [21] or a clustering model [23]. Using the low-dimensional parameter space, compared to the slider-based system [24], our approach requires a considerably smaller number of parameters for emotion controls.

5. Conclusion

In this paper, we presented a real-time approach to create an emotional speech animation with emotion control. In our approach, an example-based scheme was proposed to generate a 3D lip-synced animation with coarticulation effects and emotional expressions from a small set of example models based on the multidimensional scattered data interpolation. To generate convincing animation results, an importance-based scheme was introduced to combine the viseme and expression models into the final animation sequence during the animation generation. As demonstrated in the experimental results, our approach can be applied to digital content or applications that use a virtual character showing a facial animation.

Currently, there are several issues for ongoing improvement. To show more details of emotional expressions, we are working to include subtle facial movements using motion capture data. As our lip-synchronization mainly relies on the phoneme length, we are considering other factors such as accents and intonation of voice, which affect lip motion. In addition, we are developing a parsing system for input text so that emotional speech animation can be generated from the emotional tags in the text in a fully automated way.

Acknowledgement

This work was supported by Institute for Information & communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No. 2020-0-00437, Proactive interaction based VRAR virtual human object creation technology) and by Basic Science Research Program through the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2017R1C1B5017000).

References

[1] C. Bregler, M. Covell, and M. Slaney, "Video rewrite: Driving visual speech with audio," in *Proc. of ACM SIGGRAPH*, pp. 353–360, 1997. Article (CrossRef Link)
[2] E. Cossato and H. Graf, "Photo-Realistic Talking Heads from Image Samples," *IEEE Transactions on Multimedia*, vol. 2, no. 3, pp. 152-163, 2000. Article (CrossRef Link)
[3] T. Ezzat and T. Poggio, "Visual Speech synthesis by morphing visemes," *International Journal of Computer Vision*, vol. 38, pp. 45-57, 2000. Article (CrossRef Link)
[4] T. Ezzat, G. Geiger, and T. Poggio, "Trainable Videorealistic Speech Animation," *ACM Transactions on Graphics*, vol. 21, no. 3, 2002. Article (CrossRef Link)
[5] M. Brand, "Voice Puppetry," in *Proc. of ACM SIGGRAPH*, pp. 21-28, 1999. Article (CrossRef Link)
[6] F. I. Parke, *A Parametric Model of Human Faces*, Ph.D. Thesis, University of Utah, 1974.
[7] A. Pearce, B. Wyvill, G. Wyvill, and D. Hill, "Speech and Expression: A Computer Solution to Face Animation," in *Proc. of Graphics Interface*, pp. 136-140, 1986. Article (CrossRef Link)
[8] G. A. Kalberer and L. V. Gool, "Lip Animation Based on Observed 3D Speech Dynamics," in *Proc. of Computer Animation 2001*, vol. 4309, pp. 20–27, 2001. Article (CrossRef Link)

[9] L. Wang, H. Chen, S. Li, and H. M. Meng, "Phoneme-level articulatory animation in pronunciation training," *Speech Communication*, vol. 54, no. 7, pp. 845-856, Sep. 2012. Article (CrossRef Link)
[10] M. Kawai, T. Iwao, D. Mima, A. Maejima, and S. Morishima, "Data-driven speech animation synthesis focusing on realistic inside of the mouth," *Journal of Information Processing*, vol. 22, no. 2, pp. 401-409, 2014. Article (CrossRef Link)
[11] F. Kuhnke and J. Ostermann, "Visual speech synthesis from 3D mesh sequences driven by combined speech features," in *Proc. of IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1075-1080, 2017. Article (CrossRef Link)
[12] Y. Lee, D. Terzopoulos, and K. Waters, "Realistic Modeling for Facial Animation," in *Proc. of ACM SIGGRAPH*, pp. 55-62, 1995. Article (CrossRef Link)
[13] E. Sifakis, I. Neverov, and Ronald Fedkiw, "Automatic determination of facial muscle activations from sparse motion capture marker data," *ACM Transactions on Graphics*, vol. 24, no. 3, pp. 417–425, 2005. Article (CrossRef Link)
[14] M. Cong, M. Bao, J. L. E, K. S Bhat, and R. Fedkiw, "Fully automatic generation of anatomical face simulation models," in *Proc. of ACM SIGGRAPH/Eurographics Symposium on Computer Animation*, pp. 175-183, 2015. Article (CrossRef Link)
[15] A.-E. Ichim, P. Kadlechk, L. Kavan, and M. Pauly, "Phace: Physics-based Face Modeling and Animation," *ACM Transactions on Graphics*, vol. 36, no. 4, pp. 1-14, 2017. Article (CrossRef Link)
[16] F. Pighin, J. Hecker, D. Lischinski, R. Szeliski, and D. H. Salesin, "Synthesizing Realistic Facial Expressions from Photographs," in *Proc. of ACM SIGGRAPH*, pp. 75-84, 1998. Article (CrossRef Link)
[17] B. Guenter, C. Grimm, D. Wood, H. Malvar, and F. Pighin, "Making Faces," in *Proc. of ACM SIGGRAPH*, pp. 55-66, 1998. Article (CrossRef Link)
[18] E. Ju and J. Lee, "Expressive Facial Gestures from Motion Capture Data," *Computer Graphics Forum*, vol. 27, no. 2, pp. 381-388, 2008. Article (CrossRef Link)
[19] M. Lau, J. Chai, Y.-Q. Xu, and H.-Y. Shum, "Face poser: Interactive modeling of 3D facial expressions using facial priors," *ACM Trans. on Graphics*, vol 29, no. 1, pp. 1-17, 2009. Article (CrossRef Link)
[20] V. Barrielle, N. Stoiber, and C. Cagniart, "BlendForces: A Dynamic Framework for Facial Animation," *Computer Graphics Forum*, vol 35, no. 2, pp. 341-352, 2016. Article (CrossRef Link)
[21] J. Jia, S. Zhang, F. Meng, Y. Wang, and L. Cai, "Emotional audiovisual speech synthesis based on PAD," *IEEE Transactions on Audio Speech and Language Processing*, vol. 19, no. 3, pp. 570-582, 2011. Article (CrossRef Link)
[22] P. Ekman and E. L. Rosenberg, *What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System*, Oxford University Press, 1997.
[23] V. Wan, R. Blokland, N. Braunschweiler, L. Chen, B. Kolluru, J. Latorre, R. Maia, B. Stenger, K. Yanagisawa, Y. Stylianou, M. Akamine, M. J. F. Gales, and R. Cipolla, "Photo-Realistic Expressive Text to Talking Head Synthesis," in *Proc. of Annual Conference of the International Speech Communication Association*, pp. 2667-2669, 2013.
[24] A. Stef, K. Perera, H. P. H. Shum, and E. S. L. Ho, "Synthesizing Expressive Facial and Speech Animation by Text-to-IPA Translation with Emotion Control," in *Proc. of International Conference on Software, Knowledge, Information Management & Applications*, pp. 1-8, 2018. Article (CrossRef Link)
[25] C. G. Fisher, "Confusions among visually perceived consonants," *Journal of Speech and Hearing Research*, vol 11, pp. 796-804, 1968. Article (CrossRef Link)
[26] J. A. Russel, "A Circomplex Model of Affect," *Journal of Personality and Social Psychology*, vol. 39, pp. 1161-1178, 1980. Article (CrossRef Link)
[27] P.-P. Sloan, C. F. Rose, and M. F. Cohen, "Shape by example," in *Proc. of Symposium on Interactive 3D Graphics*, pp. 135-144, 2001. Article (CrossRef Link)

[28] H. Pyun, Y. Kim, W. Chae, H. W. Kang, and S. Y. Shin, "An Example-Based Approach for Facial Expression Cloning," in *Proc. of Eurographics/SIGGRAPH Symposium on Computer Animation*, pp. 167-176, 2003. Article (CrossRef Link)
[29] VoiceText. Available online: http://www.voiceware.co.kr (accessed on 1 June 2019).

Wonseok Chae received his BS and MS degree in computer science from the Korea Advanced Institute of Science and Technology, Daejeon, Rep. of Korea, in 2001 and 2003, respectively. From 2003 to 2005, he worked at the Digital Printing Division, Samsung Electronics Co. Ltd., Suwon, Rep. of Korea, as an engineer. He has been a senior researcher at the Creative Content Research Division, ETRI, Daejeon, Rep. of Korea, since 2005. His research interests include computer graphics techniques in VFX, virtual reality, and augmented reality.

Yejin Kim received his BS degree in computer engineering from the University of Michigan, Ann Arbor, USA, in 2000. He received his MS and PhD degrees in computer science from the Korea Advanced Institute of Science and Technology, Daejeon, Rep. of Korea, in 2003 and the University of California, Davis, USA, in 2013, respectively. From 2003 to 2013, he worked at the Visual Contents Research Department, ETRI, Daejeon, Rep. of Korea, as a research scientist. Currently, he is an assistant professor at Hongik University, Sejong, Rep. of Korea. His research interests include 3D character animation and authoring techniques in computer graphics.