ROMANIAN JOURNAL OF INFORMATION SCIENCE AND TECHNOLOGY
Volume **15**, Number 3, 2012, 229–245

An efficient 3D Visual Speech Synthesis Framework for Romanian Language Logopedics use

Mihai Daniel ILIE, Cristian NEGRESCU, Dumitru STANOMIR

Department of Telecommunications,
Politehnica University of Bucharest, Romania
E-mail: mihai.iliedamaschin@yahoo.com
E-mail: {negrescu, dumitru.stanomir}@elcom.pub.ro

Abstract. In this paper, we propose a 3D facial animation model for simulating visual speech production in the Romanian language. To our best knowledge, this is the first study dedicated to 3D visual speech synthesis for the Romanian language. The model is also capable to generate complex emotion activity for the virtual actors during the speech. Using a set of existing 3D key shapes representing various facial expressions, fluid animations describing facial activity during speech pronunciation are provided, taking into account several Romanian language coarticulation effects. A novel mathematical model for defining efficient viseme coarticulation functions is provided. The 3D tongue activity could be closely observed in real-time while different words are pronounced in Romanian. The framework proposed in this paper is designed for helping people with hearing disabilities learn how to articulate correctly in Romanian and also it works as a training-assistant for Romanian language lip-reading. The efficiency of our method was successfully proven by tests made on various logopedics specialists, deaf students and deaf lip-reading instructors from a special school for people with hearing disabilities in Bucharest, Romania.

Key-words: visual speech synthesis, Romanian language, logopedics, lip-reading, viseme animation, 3D facial animation, coarticulation.

1. Introduction

Since the year 1972, when Parke proposed the first virtual facial animation model [1], 3D computer facial animation has been an increasingly active research area,

spanning through several research areas from medical simulations and visual speech production and analysis applications, to the ever-growing game industry, multimedia, and cinematography. Even though several continuously improved facial animation techniques have been developed so far [2], making a 3D virtual representation of a human head behave realistically is a very difficult task, since the human eye, being used to everyday interaction with other humans, is most trained to detect even the tiniest inadvertence in a computer generated facial animation sequence. In most cases though, it would be difficult for the observer to pinpoint the exact reason why the animation fails to be credible.

Speech production is fundamentally related to the activity of specific organs in the vocal tract. Each particular phoneme implies a specific facial expression and a tongue position. The term "viseme" [3] refers to a cardinal concept in 3D visual speech production and it represents a visual phoneme, *i.e.* the facial expression corresponding to the pronunciation of a phoneme. As concluded by many studies such as [4], the lips position while pronouncing a specific phoneme is context dependent, which means the lips aspect for the same phoneme differs depending on the speech segments located immediately before and after the current viseme. This effect is called coarticulation. The quality and realism of any visual speech animation highly depends on the way coarticulation is modeled. Coarticulation may be anticipatory, when a certain viseme is influenced by a following one, or preservative, in which case the current viseme lies under the influence of a preceding one. Coarticulation effects are also language specific, as shown in [5].

The mainstream method for animating 3D faces is the multiple morph target one, also known as the blendshape method [6]. The 3D head is regarded as a discrete domain mesh M = (V, T), where V is a finite set of vertices in 3D space and T is the set of triangles that describe the way vertices are linked to one another. By manually displacing the object's vertices, several instances of the initial head are obtained, each representing a different facial expression of the virtual actor and all sharing the same topology. Such a deformed instance of the initial object is called a blendshape item or morph target. The animation is performed by interpolating the 3D head surface between several morph target combinations. Providing the blendshapes are sufficiently numerous, the blendshape weights are carefully chosen throughout time and the morph targets are anatomically correctly sculpted, the resulting 3D facial animation is quite convincing.

In this paper, we propose a method for animating 3D virtual faces in order to best describe lip movements and tongue articulation processes for the pronunciation of any phrase or single word in Romanian. For this purpose, we use the blendshape method on different virtual head actors, by applying our proposed visual speech coarticulation model for Romanian language.

We also propose a model for incorporating complex emotion activity in the speech, including eye-blinking and other human natural-acting effects.

To our knowledge, no 3D visual speech animation framework designed for the Romanian language has been developed so far.

The set of particular viseme blendshapes required for correct Romanian language pronunciation facial expressions have been manually modeled by the author of this

paper, under the close and careful assistance of logopedics specialists from the School for people with hearing disabilities *Sfânta Maria*, from Bucharest, Romania.

The implementation of our application was done using the QT IDE for C++ and the common OpenGL libraries.

Also, in this paper, the Romanian-specific phonetic groups and Romanian words were written together with their IPA (International Phonetic Alphabet) equivalent placed between square brackets, when needed. This is particularly important for people who do not speak Romanian.

2. Related Work

The most commonly featured synthetic visual speech coarticulation model is the Cohen-Massaro one [7], which proposes the use of dominance functions to describe the influence of each visual speech segment over the surrounding ones by modeling morph weight values over time. The dominance function proposed for a speech segment s is a negative exponential function:

$$D_s = \alpha_s \cdot e^{(-\theta^{\pm} \cdot |\tau|^c)}, \tag{1}$$

where τ is the time offset relative to the dominance peak, α is the magnitude, θ^+ and θ^- represent the anticipatory and preservative rate parameters, and c is an exponent coefficient. The animation is obtained by assigning different parameters to the functions associated to each speech segment, depending on the relation between these segments. This method provides overly smoothed curves for viseme animation, making it impossible for some viseme blendshapes to reach unitary weights. Goyal et al. [8] improve the method by adding an eye-blink linear model in order to enhance the realism of the animation.

Ma et al. [9] propose the use of several visual coarticulation rules available for English, together with a kernel smoothing technique using a superquadric curve as a kernel function:

$$K(u) = C(1 - |u|^{\alpha})^{\frac{1}{\alpha}}, \tag{2}$$

where C is a kernel constant value and α is the exponent coefficient. This method provides very good results for the English language. Other consistent results have been obtained by Huang et al. [10], by using weighted Gaussian dominance functions for the vowel influence over surrounding visual speech segments. In Huang's work, the dominance function is described as:

$$D_j(t_j) = exp\left(\frac{t_j - \mu_j}{(d_j \times \sigma_j)^2 + \varepsilon}\right), \tag{3}$$

where μ_j is the center time of the current viseme, d_j is its duration and σ_j is a constant specific to the current viseme, which describes its influence on the surrounding ones.

Significant researches have been undergone for other languages as well, such as the one of De Martinoa et al. for Brazilian Portuguese [11] or Bothe et al. [12] and Albrecht et al. [13] for the German language.

Wang et al. [14] propose a 3D facial animation framework based on the Cohen-Massaro model which also provides emotional activity during speech production. All these methods are based on the use of coarticulation rules available for specific languages and also on raw video recorded material capable of providing useful information regarding lips movement. This latter task is done by measuring the geometrical displacements of specific mouth key points when comparing different frames of a video-recorded speech performance from a real human.

3. A 3D Visual Speech Production Model for the Romanian Language

For the model proposed in this paper, we have manually modeled the necessary Romanian specific visemes as blendshapes for several 3D head test objects, using observational data from deaf lip-reading instructors and the close assistance of the logopedics specialists from a school for people with hearing disabilities. We have also used valuable speech production information regarding tongue position, teeth and maxillary positions for the pronunciation of each phoneme in Romanian, which are exhaustively detailed in [15]. In our work, we use a number of 25 visemes for Romanian visual speech synthesis, for each one of the various virtual actors we modeled. Figure 1 shows our test virtual actors.

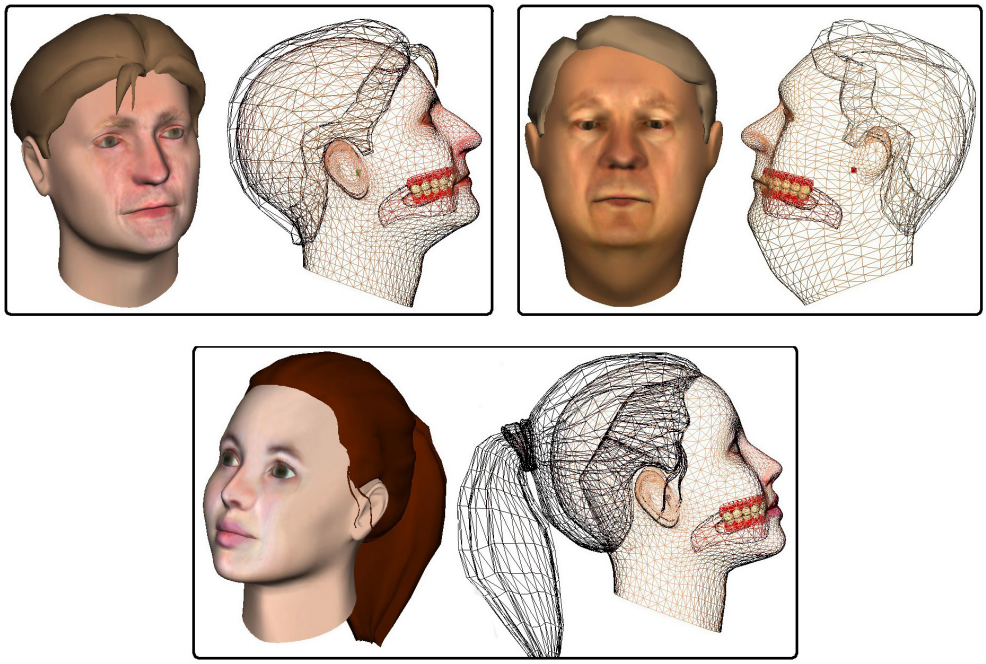

Fig. 1. Test 3D virtual actors used in our work.

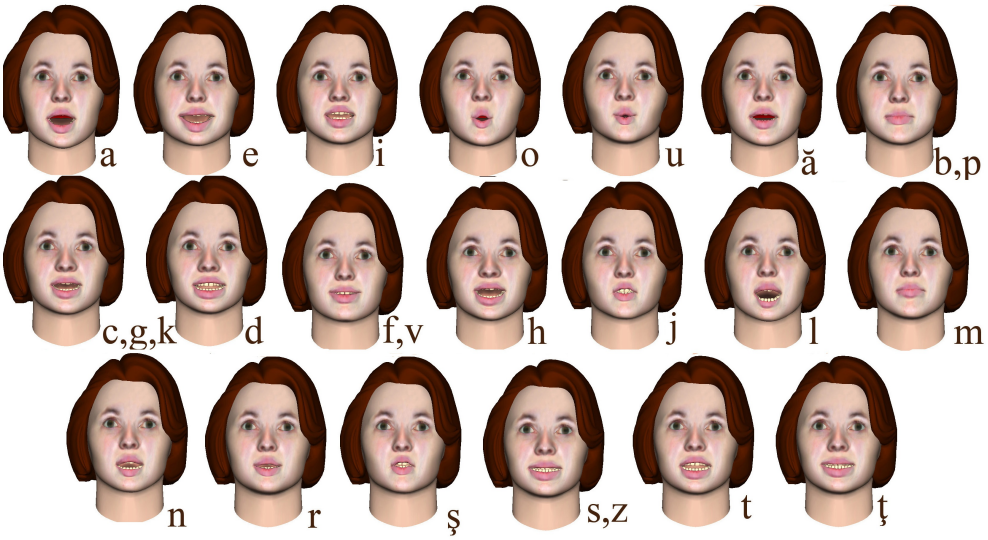

Fig. 2. Romanian language visemes modeled for one of the actors from Figure 1.

Figure 2 and Figure 3 show various visemes we modeled according to detailed Romanian lip-reading reference data [15] and explicit viseme images from the Romanian Abecedary for deaf people [16]. These blendshapes have also been validated by several deaf lip-reading instructors from a school for deaf people in Bucharest.

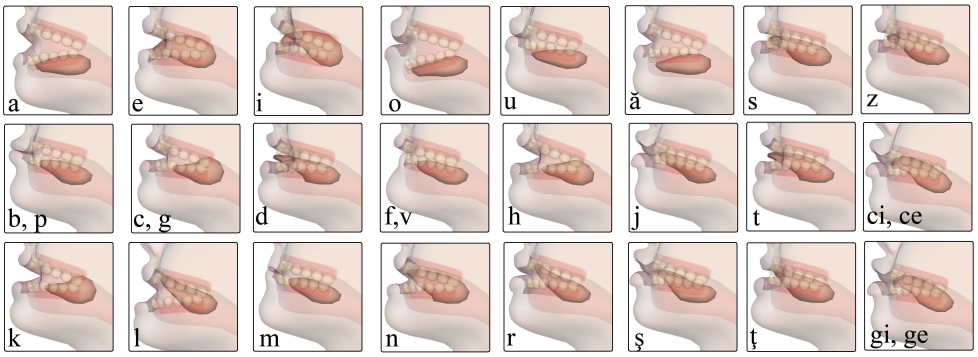

Fig. 3. Tongue positions we use for various visemes, modeled using detailed reference from [15].

Also, in order to add emotional activity to the visual discourse, we modeled six blendshapes for the universal facial expressions, according to [17]. These universal emotions are happiness, sadness, anger, disgust, fear and surprise. Figure 4 shows the associated basic emotions corresponding to one of the actors from Figure 1.

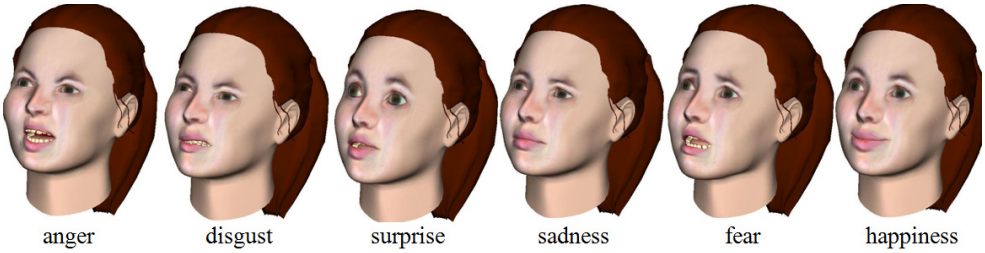

Fig. 4. The six universal facial expressions used in our work to model emotional activity.

The resulting blendshape deformed object at an arbitrary time moment t during the speech is:

$$S(t) = S_0 + \sum_{i=1}^{N_s} w_i(t) \cdot (S_i - S_0), \tag{4}$$

where S_0 denotes the neutral pose object, N_s is the total number of blendshapes (viseme blendshapes and also emotion blendshapes), S_i refers to the i indexed blendshape and $w_i(t)$ is a weight value function of time specific to each S_i morph target. To avoid abnormal results, for any moment t the following condition is generally respected:

$$\sum_{i=1}^{N_s} w_i(t) \leq 1. \tag{5}$$

The Romanian phonetic groups *ce* [t͡ʃe], *ci* [t͡ʃi], *ge* [d͡ʒe], *gi* [d͡ʒi], *che* [ce], *chi* [ci], *ghe* [ɟe], *ghi* [ɟi] each require an entire sequence of visemes in order to be properly described, as presented in Figure 5.

Fig. 5. Romanian-specific phonetic groups we use for our visual speech simulation model.

The *x* consonant in Romanian is best described as an adjointment of two consonants: *c* and *s* in some cases, and *g* and *z* in others.

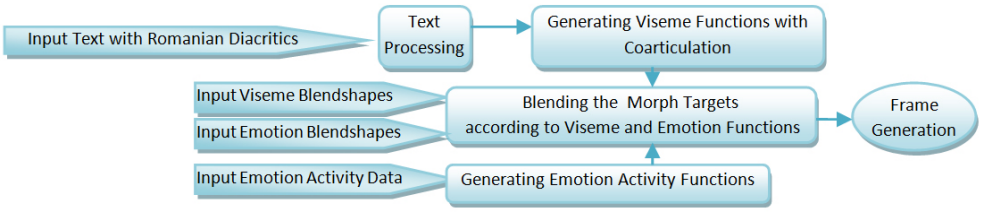

Fig. 6. Framework of our visual speech animation model.

The general framework of our speech production model is described in Figure 6. All blendshapes (including visemes and emotions) are blended together additively using equation (4). The input text is first automatically processed using the general syllable separation rules for Romanian language available in the Ortographic and Orthoepic Romanian Dictionary [18]. This is required in order to identify diphthongs, triphthongs and hiatus structures in Romanian and therefore determine the vowels, semivowels and consonants of the text. Two linked lists are then generated for each viseme in particular. The first list stores four chronological time moments for each viseme occurrence: the moment it starts, the moment its maximum magnitude is reached, the moment it begins to lose magnitude and eventually the moment its influence returns to 0. The other list stores the maximum magnitude for each occurrence of this viseme. We shall refer to these lists as the viseme occurrence list and the viseme maximum magnitude list. The elements of these lists serve as control knots for later generation of the viseme dominance functions.

The decisions for choosing time moments and magnitudes for each viseme are based on two major criteria. The first criteria regards empirical observation data we have extracted from video recordings of real human actors pronouncing various words in Romanian. The geometrical displacements of key geometrical features of the mouth area are observed, thus gaining information regarding the duration units of coarticulation effects in Romanian and also for the maximum magnitude for each viseme morph target in different phonetic contexts.

The second criteria deals with several Romanian-specific coarticulation effects [18]. For exception cases of different words that are written in the same way, accents are required in order for the algorithm to decide upon the correct syllable-separation. Such is the case of the word *hainǎ*, which has different meanings and pronunciations as an adjective [ha.i.nə] and as a noun [ˈhaĭ.nə]. Therefore, the correct spelling for its adjective meaning shall have to be *haínǎ*, in which case *í* is a vowel. Also, *i* is exceptionally treated as a semivowel [ĭ] for the case of plural nouns ending in a single *i* (such as *plopi*, *stupi*, *mori*).

All vowel visemes span a considerable influence over their surrounding visual speech segments, be them visemes associated to semivowels or consonants. Generally, consonants are affected in various amounts by preceding or following vowels. The only exceptions are made by the labial consonants p, b, m, f and v, which need to be less influenced by the vowels in order for the lips to close and thus ensure a correct movement. Nevertheless, slight influences such as lips rounding before u or o still happen. Also, lip shapes during pronouncing semivowels are affected by the lip

shape of their neighboring vowel. In some phonetic contexts, a vowel viseme shape may span its influence over more speech segments than only its direct neighbors. For instance, in the case of the Romanian word *croitor* ['kro.i.tor], the vowel *o* imposes the lip rounding aspect on both *c* and *r* consonants (but in different amounts). The same happens for the Romanian word *ciob* ['t͡ʃiob], in this case for the *ci* phonetic group affected by the following vowel. Some of the dental consonants in Romanian language (such as *t*, *d*, *s* and *ţ* [t͡s]) are also slightly affected in an anticipatory coarticulation effect by following consonants, when these latter consonants are labial ones. A relevant example is the Romanian word *astm*. In such situations, the viseme maximum magnitude associated to consonants such as the *t* here are severely reduced in order to ensure smooth transitions.

Another important aspect regards tongue activity during pronunciation. Even though lip shapes during pronouncing consonants are affected by their neighboring speech segments, the tongue positions are not. For instance, in the case of *t*, even if the lips' shape is diminished due to influence from a following vowel, the tongue still has to touch the back of the upper teeth in order for a correct sound to be pronounced. Therefore, in our work the tongue is animated separately from the lips and is assigned a different set of animation parameters. Its maximum magnitude values are always equal to 1.

To grasp the "four time moments" for each viseme occurrence, we undergo the following algorithm. The speech-text is stored in a string of characters. We use two itinerant time moment variables denoted as $last_1$ and $last_2$. These two variables are assigned the initial values 0 seconds and 0.25 seconds and always show the last two time moments of the previous viseme. The algorithm loops through the character string and for each character executes the following instructions:

$$
\begin{gathered}
t_1 \leftarrow last_1 \\
t_2 \leftarrow last_2 \\
t_3 = t_2 + q_1 \\
t_4 = t_3 + q_2 \\
\alpha = q_3 \\
last_1 = f(t_1, t_2, t_3, t_4) \\
last_2 = g(t_1, t_2, t_3, t_4)
\end{gathered}
$$

The four time moments associated to the viseme occurrence implied by the current string character are determined as shown in the pseudo-code algorithm above. Also, for each viseme occurrence, its specific maximum magnitude value (denoted as α in the pseudo-code above) is computed. The q_1, q_2 and q_3 quantities and the two functions $f(t_1, t_2, t_3, t_4)$ and $g(t_1, t_2, t_3, t_4)$ differ depending on the phonetic context of the current phoneme (its coarticulation relation to the previous and next speech segments). They are particular to each consonant, vowel and semivowel for each one of its possible phonetic contexts, and were chosen using the two major criteria previously presented in this section. Generally, due to speech fluency, the distance between t_2 and t_3 is very short (between 0.02 and 0.05 seconds). $last_1$ and $last_2$ are re-computed so as to provide start-moment information for the next viseme. If no coarticulation

effects happen, it means the two functions have the particular forms $f(t_1, t_2, t_3, t_4) = t_3$ and $g(t_1, t_2, t_3, t_4) = t_4$. This generally happens for hiatus structures (vowel–vowel groups). In other cases (consonant–vowel, vowel–consonant, vowel–semivowel and semivowel–vowel groups), the two functions take more complex forms (linear combinations of t_1, t_2, t_3, t_4) based on the two major criteria previously presented in this section. For instance, if the current character is the semivowel *o* [o̯] and the next one is the vowel *a*, we choose $f(t_1, t_2, t_3, t_4) = 0.5 \cdot t_1 + 0.5 \cdot t_2$ and $g(t_1, t_2, t_3, t_4) = t_4 + 0.1$ in order to model the influence of the vowel over its preceding semivowel. This is the general form we use for all semivowel vowel diphthongs that include the vowel *a*.

Therefore, at the end of this processing part, two linked lists are obtained for each viseme. The viseme dominance functions are constructed using these arrays and are used in equation (4) as weight functions of time. An arbitrary element of the viseme occurrence list associated to a S_i viseme blendshape is denoted as t_{ijk}, with $j = \overline{1, n_i}$ and $k = \overline{1, 4}$. n_i is the total number of apparitions of the S_i viseme in the speech and k describes to which one of the four key moments of the j–th viseme apparition t_{ijk} refers to. The other list stores maxim magnitude values for each such viseme apparition. Each group $(t_{ij1}, t_{ij2}, t_{ij3}, t_{ij4})$ has its corresponding α_{ij} value from the viseme maximum magnitude list. Considering the j indexed appearance of the S_i viseme, we propose the following expression for a viseme function associated to S_i:

$$V_i(t) = \begin{cases} \alpha_{ij} \cdot \left[(1 - \tau_1^{\omega_1}) \cdot \left(1 - \cos\left(\frac{\pi}{2} \cdot \tau_1\right)\right) + \tau_1^{\omega_1} \cdot \sin\left(\frac{\pi}{2} \cdot \tau_1\right)\right] & \text{if } \ t \in [t_{ij1}, t_{ij2}] \\ \alpha_{ij} \quad \text{if } \ t \in [t_{ij2}, t_{ij3}] \\ \alpha_{ij} \cdot \left[\tau_2^{\omega_2} \cdot \left(1 - \cos\left(\frac{\pi}{2} \cdot (1 - \tau_2)\right)\right) + \left(1 - \tau_2^{\omega_2}\right) \cdot \sin\left(\frac{\pi}{2} \cdot (1 - \tau_2)\right)\right] \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \textit{if } \ t \in [t_{ij3}, t_{ij4}] \end{cases} \tag{6}$$

The notations τ_1 and τ_2 are:

$$\tau_1 = \frac{t - t_{ij1}}{t_{ij2} - t_{ij1}} \quad , \quad \tau_2 = \frac{t - t_{ij3}}{t_{ij4} - t_{ij3}} \tag{7}$$

and ω_1 and ω_2 are exponent coefficients used to model the smoothness or abruptness of the ascending and descending parts of the curve for a particular viseme appearance. Generally, the chosen values for these coefficients are close to 1, yet there are cases where other values are recommended. For example, for a vowel considerably affecting the surrounding visual speech segments, values between 0 and 1 are required. Higher than 1 values will result into narrowing of the viseme appearance peaks.

Figure 7 shows different possible aspects of the two parts of a viseme occurrence curve for a j-indexed appearance.

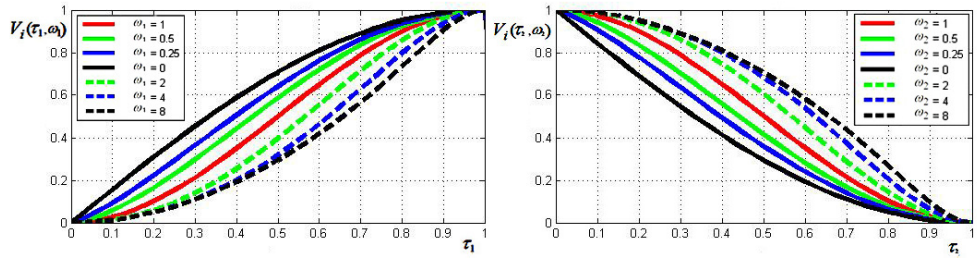

Fig. 7. Various possible aspects for different ω_1 and ω_2 values.

4. Integrating Emotional Activity

For the emotion activity input data, we employ the following approach. The moment when one certain emotion starts or ends during the speech is specified by the user by adding a tag with the emotion's name written in Romanian between left and right angled brackets, while introducing the speech text. We use a number of six tags, associated to each of the six universal emotions. Based on the emotion tags spread through the speech text, a linked list of emotion appearance values is built, similar to the viseme appearance arrays, implying four necessary key moments for each emotion occurrence. The start and end moments for an emotion occurrence are determined by the moments its two tags are placed in the speech. Researches regarding human facial activity [17] show that generally emotion changes are sudden, which means emotion-related facial expressions change very fast from one to another during the speech. That is why we choose rather short intervals (between 0.5 and 0.75 seconds) for the raise and fall of the curve associated to the emotion occurrences. The emotion weight functions are determined using the same mathematical model we use for the viseme functions, *i.e.* equations (6) and (7). The emotion activity is then summed up with the speech activity using equation (4). For instance, for an input phrase such as *<trist>Azi a plouat <trist> <vesel> şi apoi a fost soare <vesel>*, the virtual actor utters the first sentence with a sad look on her face and the second one with a happy facial expression.

The arising problem though is that the mouth and maxillary expression of an emotion could severely distort the lips position during normal speech if the two blendshapes are added together. Therefore, only the upper part of the face has to be affected by the emotion activity functions. For this, we use the automatic masked morphing algorithm we presented in [19] in order to define specific masks for each emotion in particular. First, a raw selection of the upper part of the head is defined. The algorithm automatically extends the mask selection in those areas where the two objects (the emotion one and the speech one) are most dissimilar, by allowing a gradual and smoothly decreasing influence around the mask. As described in [19], the dissimilarity level for various regions around the selection border is measured using the Euclidian distance between the initial and target positions associated to each vertex on the border. Figure 8 shows the example of a fear/scared emotion which is

added to the speech flow, by using the algorithm presented in [19]. The emotion mask grows larger only for the cheek area of the character, since the emotion blendshape and the neutral speech blendshape are more dissimilar in that particular region.

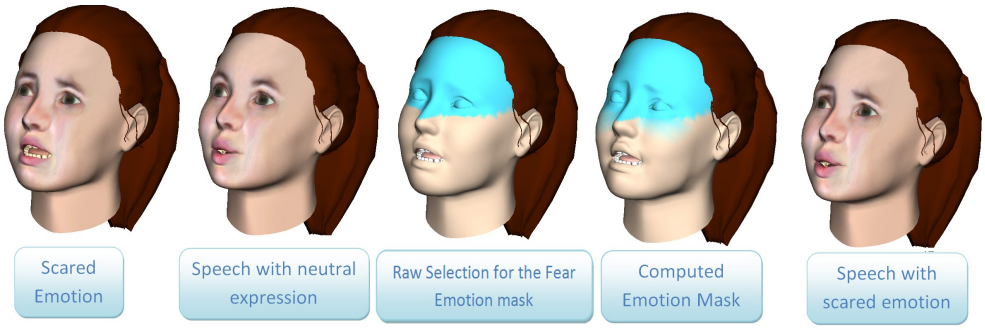

Fig. 8. Adding a fear/scared emotion to a neutral expression speaking virtual 3D actor.

Our facial animation speech model also performs emotional activity changes depending on the orthographical signs present in the text. For this, we make use of specific blendshapes which are applied locally for the arcades or for the mouth corner areas. For example, the interrogative sign *?* implies gradually raising the eyebrows towards the end of the sentence, whereas the exclamation sign *!* also induces a slight widening of the mouth during speech. These effects add realism to the visual speech animation flow. Another aspect regarding virtual 3D actor performance during speech regards eye blinking. An actor that speaks with no blinks fails to be credible to any observer, even though it would be difficult for one to grasp the exact reason why the animation is not realistic. The blinking activity is also performed using the blendshape animation method, *i.e.* equations (4) and (6). Based on empirically observed data regarding real human habitual gestures, we concluded that a normal eye blink lasts very short (less than one second), so we choose a value of 0.75 seconds for each blink and add an occurrence condition: the distance between two consecutive blinks has to vary between 1 and 5 seconds. The blink occurrence times are chosen randomly, respecting these thresholds.

5. Animating Visual Speech for Non-human Characters

In equation (4), we have employed the linear interpolation approach for morphing between various blendshapes. This brings about very good results for virtual human 3D objects. At all events, in the case of some non-human actors, such as animals, the metamorphosis between two different composures implies large scale rotations or considerable deformations, in which case linear interpolation may cause object shrinking or object self interaction issues. To avoid that, our facial animation framework also allows the use of an efficient non-linear interpolation method in the case of animal animations, or any other non-human characters for multimedia use. We use the non-linear interpolation method that we proposed and thoroughly described in [20]. The

method ensures fast computational costs and volume preserving during the transformation, by studying the local proprieties of each particular patch of the object's geometry and determining a different circle arc for each such patch. Basically, for each vertex, modified vertex normals are computed for both initial and target states, taking into account several local geometry features such as the variation of triangle normals, triangle areas or dihedral angles. Figure 9 presents the example of a wolf animation using our framework with both linear and circular interpolation approaches. As observed from the comparison in the right part of the figure, the interpolation method we proposed in [20] ensures volume preservation for the wolf's jaw, thus saving animation realism. The figure presents the transition between visemes m and a.

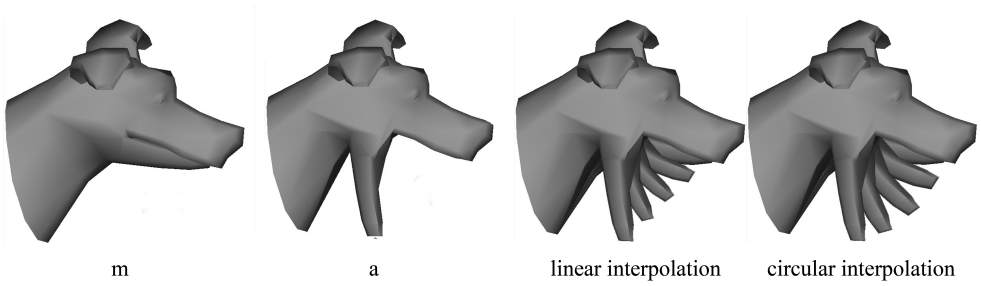

Fig. 9. Wolf speech animation using both linear and non linear interpolation methods.

6. Results and Conclusions

We have applied our visual speech animation system on various test 3D head models. Our animation model has proven its efficiency through several Romanian pronunciation tests, implying different coarticulation effects. For instance, Figure 10 presents snapshots from the animation sequences of the Romanian words *caşalot* ['ca.ʃa.lot] and *leoaică* [le'o̯aĭ.cə] pronounced using our method and the 3D test head from Figure 2. As seen in the figure, the lip aspects while pronouncing o as a vowel in *caşalot* and as a semivowel in *leoaică* are most different. In the latter case, the semivowel o implies a lip shape which looks mostly like a slight u. Also, the duration and magnitude differ in the two situations. As seen above, the Romanian consonant implies lips puckering, while the consonant l needs tongue-to-teeth raising and is much affected by its following vowel in both cases. For this reason, as seen in Figure 10, due to the anticipatory coarticulation effect the lips position differs for the two l occurrences. Figures 11 and 12 show plots of the viseme dominance functions our algorithm produces for the two words *caşalot* ['ca.ʃa.lot] and *leoaică* [le'o̯aĭ.cə] in order to generate the visual speech animation flow, as snapshot in Figure 10.

Our synthetic visual speech production model permits total user interaction with the virtual actor. The 3D head could be rotated, translated or zoomed in real-time while pronouncing different words in Romanian. Another key feature of our application regards tongue activity. By allowing different transparency degrees for

the head, teeth and oral cavity, the tongue activity could be closely observed from any angle and distance, using any user-desired playback speed.

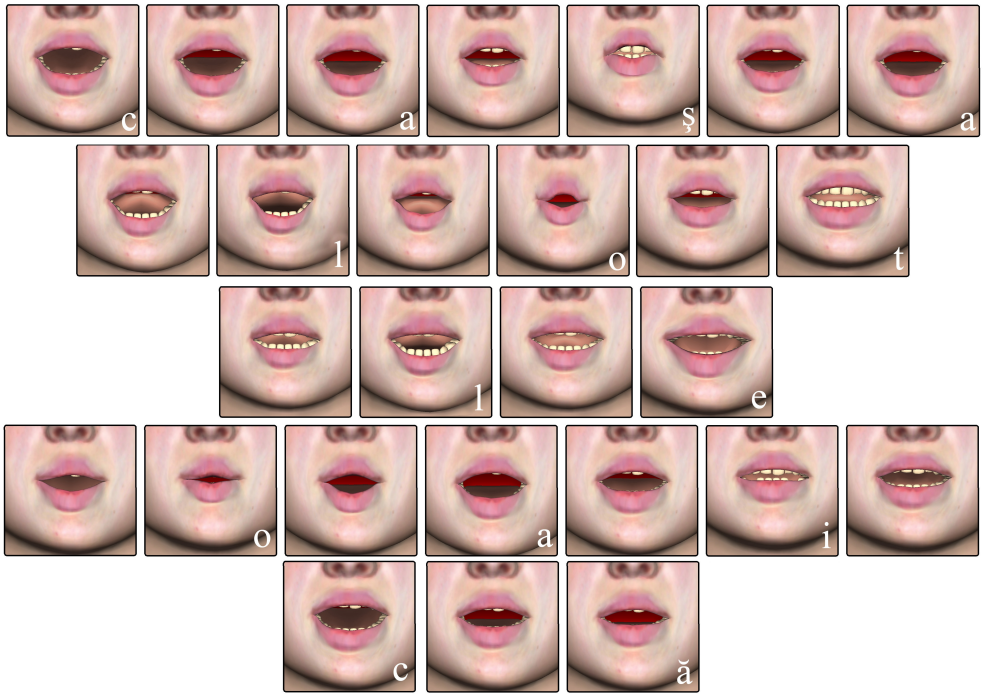

Fig. 10. Visual speech snapshots of our virtual actor for the words *caşalot* ['ca.ʃa.lot] and *leoaică* [le'o̯aĭ.cə].

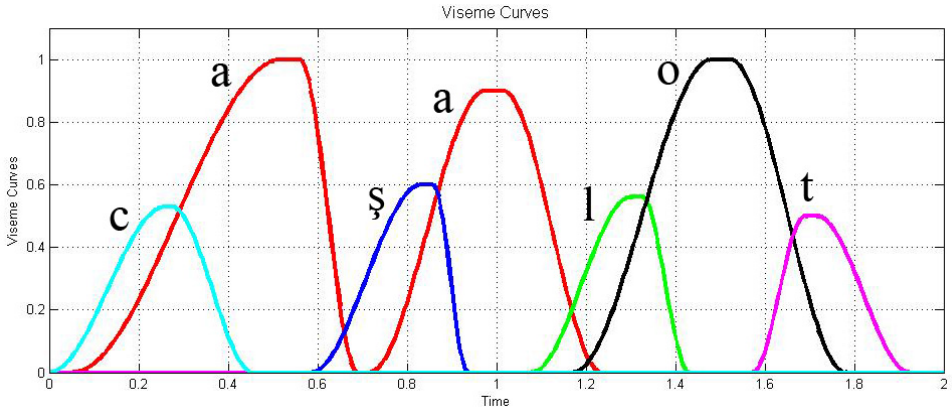

Fig. 11. Viseme Dominance Curves for the word *caşalot* from Figure 10.

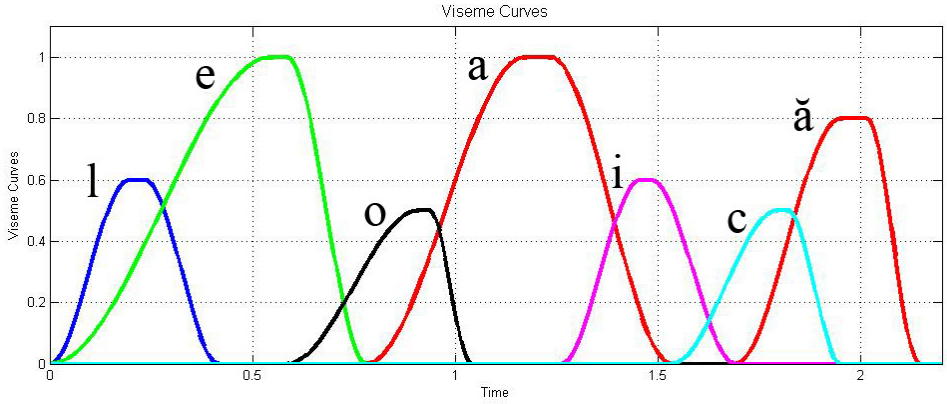

Fig. 12. Viseme Dominance Curves for the word *leoaică* from Figure 10.

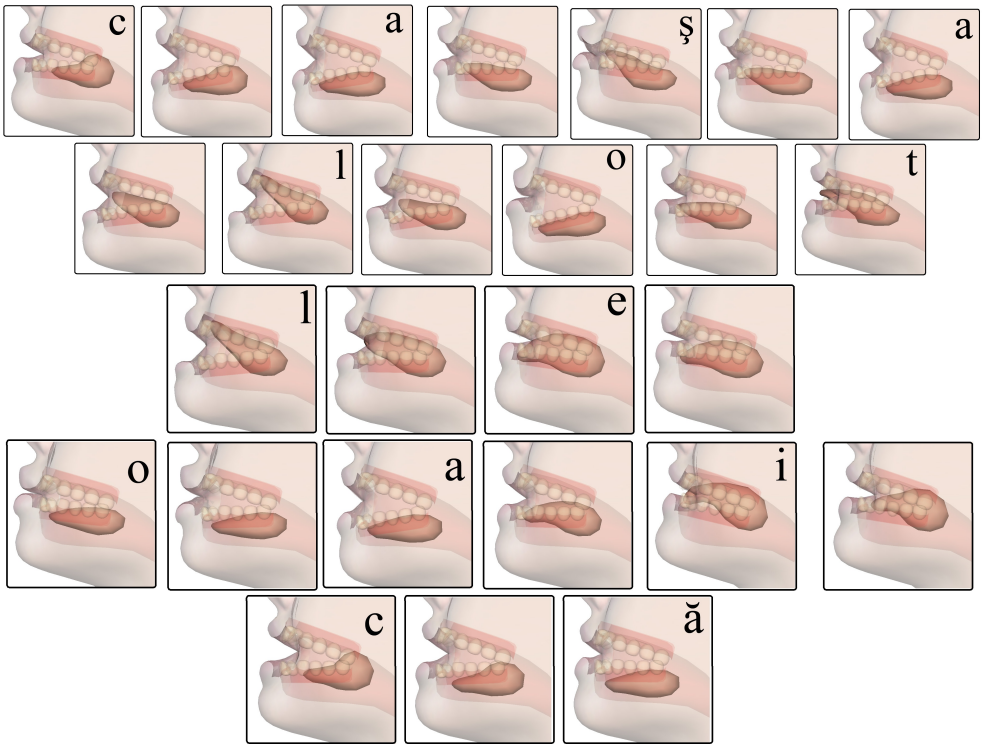

Fig. 13. Tongue activity while pronouncing the Romanian words *caşalot* ['ca.ʃa.lot] and *leoaică* [le'o̯aĭ.cə].

Figure 13 shows tongue animation snapshots related to the virtual actor from Figure 2 while pronouncing the two Romanian words from Figure 10. As seen in the picture, even if the lip shape for *l* differs when it is situated before *e* and *o*, the tongue position is quite similar for the two situations.

Figure 14 presents an example of how the same consonant cand produce different lip aspects depending on the speech segment following after it. For the Romanian expression *român ortodox* [roˈmɨn ˈor.to.doks], the consonant *r* is first follwoed by a vowel which spans influence over it, and then by another consonant.

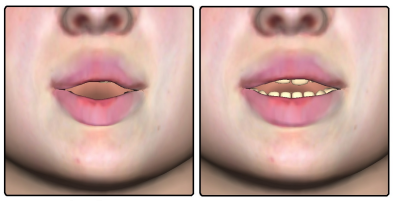

Fig. 14. Different aspects for the consonant *r* while saying *român ortodox*.

Several lip-reading tests were performed at the Special School *Sfânta Maria* for people with hearing disabilities from Bucharest, Romania, using different virtual actors. The Romanian language visual coarticulation effects and also the speech correctness from a biomechanical, anatomical and most of all logopedical point of view were validated by the logopedics specialists from the above named school and also by the deaf lip-reading instructors that teach there. Since one's ability to read lips strongly depends on one's sagacity, concentration power, IQ, cultural background, vocabulary, emotionality, general psychological state, health, memory and experience in the field, the results vary from one person to another. We therefore validated the results by comparing lip-reading tests performed on our virtual 3D actors with the ones performed on real human speakers.

Another fact of cardinal importance regarding lip-reading is that results differ considerably if the speaking actor is observed at first sight or not. That is why we compared lip-reading results of our 3D virtual actors with those of persons observed for the first time by the lip-reading instructor-specialists who were kind enough to help us and undergo the tests. After a few training hours using our facial animation framework and a single virtual actor, lip-reading results increased significantly. We therefore set the premises of our tests in such a way that the results would reflect solely the efficiency of our visual speech simulation application and not the ability and skillfulness of the deaf lip-reading instructors, which is not the subject of this paper. As discussed before, each viseme features were carefully modeled using the detailed reference [15,16] suggested by the specialists from this school and with their continuous assistance.

The set of test Romanian words was also suggested by the logopedics specialists. Most commonly used words were recommended, such as *casă* [ˈca.sə], *mama* [ma.ma], *lopată*[loˈpa.tə], *floare* [flo̯a.re], *familie* [faˈmi.lĭe], *săpun* [sə.pun], *ghem*[ˈɟem], *zăpadă*[zəˈpa.də], *cerc* [t͡ʃerk], *ghete* [ˈɟe.te], *televizor* [ˈte.le.vi.zor], *pom* [pom], *pat* [pat], *masă* [ˈma.sə], *scaun* [ˈsca.un], *peşte* [ˈpeʃ.te], *ţap* [t͡sap], *cireşe* [ˈt͡ʃi.re.ʃe], *dulap* [duˈlap] etc. Our visual speech simulation framework also allows introducing the speech-text in password echo-mode, so as for the words not to be identified by other observers.

The deaf lip-reading teachers were then asked to recognize the test words and the results were compared with lip-reading tests performed on real human speakers by the same observers. Results show a total average of 87.13% recognition rate for the virtual 3D actor we used and around 85% for real human speaker tests performed on the same lip-readers. Since it is virtually impossible for a real speaker to always pronounce each phoneme in the exact same way, our virtual actor resulting rates are slightly higher. In other words, it is somewhat easier for a lip-reader to get used to a virtual actor that acts always the same than to a human one.

The application proposed in this paper was also tested by the instructors at this school while teaching pupils and produced very satisfying results, proving its efficiency in the deaf people teaching process. Lips movements could be observed from any angle and distance during pronunciation, thus proving our application to be a very useful assistant for lip-reading trainings or any other exercise required by logopedics specialists.

To our best knowledge, this is the first study dedicated to 3D visual speech synthesis for the Romanian language.

Regarding future work, we intend to improve the quality of the visual speech simulations by undergoing further studies with different deaf lip-reading people from various backgrounds and enrich the application interface according to didactical use. The aim of our project is to make the visual speech simulation application proposed in this paper available for as many special schools for deaf people in Romania as possible, thus improving the quality of teaching for people with hearing disabilities across the country.

Acknowledgements. This work has been funded by the Sectoral Operational Program Human Resources Development 2007-2013 of the Romanian Ministry of Labour, Family and Social Protection through the Financial Agreement POSDRU/107/1.5/S/76813.

The authors would also like to thank the logopedics and deaf lip-reading specialists from School nr. 2 *Sfânta Maria* from Bucharest, Romania, for their unceasing support and valuable assistence and suggestions, especially logopedics specialist Daniela Dumitrescu.

References

[1] Parke F., *Computer generated animation of faces*, *Proceedings of the ACM National Conference*, Volume **1**, pp. 451–457, 1972.

[2] Parke F., Waters K., *Computer facial animation*, AK Peters, 2008.

[3] Fisher C. G., *Confusions among visually perceived consonants*, Journal of Speech and Hearing Research, Issue **11**, No. **4**, pp. 796–804, 1968.

[4] Owens E., Blazek B., *Visemes observed by hearing-impaired and normal-hearing adult viewers*, Journal on Speech and Hearing Research, Volume 28, pp. 381–393, 1985.

[5] Jonas N., Nartey A., *Coarticulation effects on fricative consonants across languages*, Journal of Acoustical Society of America, Volume **75**, Issue 1, p. S66, 1984.

[6] Joshi P., Tien W. C., Desbrun M., Pighin F., *Learning Controls for Blend Shape Based Realistic Facial Animation*, *Proc. of the SIGGRAPH '06*, Art. 17, 2006.

[7] Cohen M. M., Massaro D. W., *Modeling Coarticulation in Synthetic Visual Speech*, Thalman N. M., Thalman D. (editors), *Models and techniques in computer animation*, Springer, Berlin Heidelberg New York, pp. 139–156, 1993.

[8] Goyal U. K., Kapoor A., Kalra P., *Text-to-Audiovisual Speech Synthesizer*, VW '00, *Proceedings of the Second International Conference on Virtual Worlds*, pp. 256–269, Springer-Verlag, London, UK, 2000.

[9] Ma J., Cole R., *Animating visible speech and facial expressions*, The Visual Computer: International Journal of Computer Graphics, Volume **20**, Issue 2, pp. 86–105, Springer-Verlag, 2004.

[10] Huang F. C., Chen Y. M., Wang T. H., Chen B. Y., Guan S. H., *Animating Lip-Sync Speech Faces by Dominated Animeme Models*, *Proceedings of the SIGGRAPH '09*, Article no.2, New Orleans, 2009.

[11] De Martinoa J. M., Magalhaesa L. P., Violaro F., *Facial animation based on context-dependent visemes*, Journal of Computers and Graphics, pp. 971–980, 2006.

[12] Bothe H. H., Rieger F., *Visual speech and coarticulation effects*, *Proceedings of the IEEE ICASSP '93*, pp. 634–637, Mineapolis, USA, 1993.

[13] Albrecht I., Haber J., Kahler K., Schroder M., Seidel H.-P., *May I talk to you?:-) Facial Animation from Text*, *PG '02, Proceedings of the 10th Pacific Conference on Computer Graphics and Applications*, pp. 77–86, Beijing, China, 2002.

[14] Wang A., Emmi M., Faloutsos P., *Assembling an Expressive Facial Animation System*, *Proceedings of the 2007 ACM SIGGRAPH Symposium on Video Games*, pp. 21–26, 2007.

[15] Manolache C. Gh., *Surdomutitatea*, Editura Medicală, Bucharest, 1980 (in Romanian).

[16] Caraman L. M., *Abecedar pentru şcolile speciale de hipoacuzici şi surzi*, Editura Didactică şi Pedagogică, Bucharest, 1980 (in Romanian).

[17] Ekman P., Friesen W., *Unmasking the Face: A Guide to Recognizing Emotions From Facial Expressions*, Malor Books, 2003.

[18] Romanian Academy, *DOOM - Dicţionarul Ortografic, Ortoepic şi Morfologic al Limbii Române*, Romanian Academy Printing House, Second Edition, 2010 (in Romanian).

[19] Ilie M. D., Negrescu C., Stanomir D. *Automatic Masked Morphing for 3D Facial Animations*, *Proceedings of the 3rd International Conference on Future Computer and Communications, ICFCC '11*, ASME Press New York, pp. 345–350, 2011.

[20] Ilie M. D., Negrescu C., Stanomir D., *Circular Interpolation for Morphing 3D Facial Animations*, Romanian Journal of Information Science and Technology, Volume **14**, Issue 2, pp. 131–148, 2011.