ROMANIAN LANGUAGE COARTICULATION MODEL FOR VISUAL SPEECH SIMULATIONS

Mihai Daniel Ilie, Cristian Negrescu, Dumitru Stanomir

Politehnica University of Bucharest, Department of Telecommunications

ABSTRACT

In this paper, we propose a 3D facial animation model for simulating visual speech production in the Romanian language. Using a set of existing 3D key shapes representing facial animation visemes, fluid animations describing facial activity during speech pronunciation are provided, taking into account the Romanian language coarticulation effects which are discussed in this paper. A novel mathematical model for defining efficient viseme coarticulation functions for 3D facial animations is also provided. The 3D tongue activity could be closely observed in real-time while different words are pronounced in Romanian, by allowing transparency for the 3D head model, thus making tongue, teeth and the entire oral cavity visible. The purpose of this work is to provide a framework 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.

Index Terms — 3D facial animation, visual speech, viseme animation, coarticulation, Romanian language

1. INTRODUCTION

3D Computer facial animation has been an increasingly active research area for the past few decades. Since the pioneering work of Parke in 1972 [1], computer facial animation has gained a cardinal place in domains such as medical simulations, psychology, visual speech production and analysis applications, game industry, multimedia, and especially in the cinematography field. 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. In some cases, the same lip expression could describe more than one phoneme, such as for *s*, *d* and *t*. Nevertheless, the tongue position is different for these three consonants.

Another key aspect is that the lips position while pronouncing a specific phoneme is context dependent. It differs depending on the visual speech segments situated before or 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 preceeding viseme. Also, coarticulation effects are language specific [4].

One of the most frequently addressed facial animation techniques is the multiple morph target one, also known as the blendshape method [5]. 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 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 objects, by applying our proposed visual speech coarticulation model for Romanian language. To our knowledge, no 3D visual speech animation framework designed for the Romanian

© EURASIP, 2012 - ISSN 2076-1465

language has been developed so far. The 3D head models used in this paper were generated using the FaceGen Software Package and the set of particular viseme blendshapes required for correct Romanian language pronunciation facial expressions have been manually modeled by the author of this paper. In this paper, the Romanian phonetic groups and words were mentioned together with their IPA (International Phonetic Alphabet) equivalent between square brackets when needed. The implementation of our application was done using the QT IDE for C++ and the common OpenGL libraries.

2. RELATED WORK

The most commonly featured synthetic visual speech coarticulation model is the Cohen-Massaro one [6], 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:

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

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.

Ma et al. [7] 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:

$$K(u) = C(1 - |u|^{\alpha})^{1/\alpha} \quad (2)$$

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. [8], by using weighted Gaussian dominance functions for the vowel influence over surrounding visual speech segments. Significant researches have been undergone for other languages as well, such as the one of De Martinoa et al. for Brazilian Portuguese [9] or Bothe et al. [10] for the German language. Wang et al. [11] 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 ROMANIAN-SPECIFIC SYNTHETIC VISUAL SPEECH PRODUCTION MODEL

In our work, we have modeled the necessary Romanian specific visemes as morph targets for several 3D head test objects, using detailed anatomical reference and observation data. The total viseme amount we use is 25. Tongue, teeth and the palatal area are also carefully modeled. Special treatment has been applied for the Romanian language particular phonetic groups *ce* [t͡ʃe], *ci* [t͡ʃi], *ge* [d͡ʒe], *gi* [d͡ʒi], *che* [ce], *chi* [ci], *ghe* [ɟe], *ghi* [ɟi]. For the first four ones, the used viseme implies lips moving forward and is immediately followed by the *i* or *e* visemes for Romanian. The last four groups also imply a sequence of two visemes closely "glued" together: the first one is somewhat similar to the *c* consonant lips configuration, only that it requires a different tongue position. The second one is one of the two vowels. Diacritics for Romanian are also assigned special viseme shapes.

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

$$S(t) = S_0 + \sum_{i=1}^{N_s} w_i(t) \cdot (S_i - S_0) \quad (3)$$

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

$$\sum_{i=1}^{N_s} w_i(t) \le 1 \quad (4)$$

The general framework of our speech production model is described in Figure1:

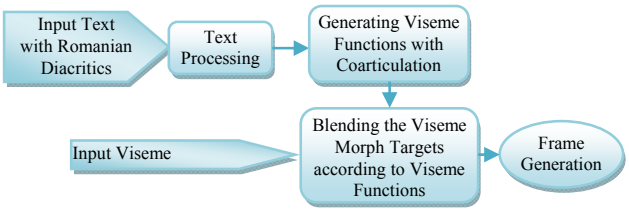

Figure1: Framework of our visual speech animation model

The input text is first automatically processed using the general syllable separation rules for Romanian language available in the Ortographic and Orthoepic Romanian Dictionary [12]. 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 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 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 coaritculation 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 [12]. First, all vowel visemes span a considerable influence over their sourrounding visual speech segments, be them visemes associated to semivowels or consonants. Generally, consonants are affected in various amounts by preceeding 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"* [ˈkro.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 word *"ciob"* [ˈt͡ʃiob], in this case for the *ci* [t͡ʃi] 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 Romaniam word *"astm"* [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 the 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 most of the time equal to 1.

Therefore, at the end of this processing part, two linked lists are obtained for each viseme. The viseme functions are constructed using these arrays and are used in equation (3) 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:

$$V_i(t) = \alpha_{ij} \cdot \left((1-\tau_1^{\omega_1}) \cdot \left(1-\cos\left(\frac{\pi}{2}\cdot\tau_1\right)\right) + \tau^{\omega_1} \cdot \sin\left(\frac{\pi}{2}\cdot\tau_1\right)\right) \quad (5)$$

$$\text{if } t \in [t_{ij1}, t_{ij2}] \text{ ,}$$

$$V_i(t) = \alpha_{ij} \quad (6)$$

$$\text{if } t \in [t_{ij2}, t_{ij3}] \text{, and}$$

$$V_i(t) = \alpha_{ij} \cdot \left(\tau_2^{\omega 2} \cdot \left(1-\cos\left(\frac{\pi}{2}\cdot(1-\tau_2)\right)\right) + (1-\tau_2^{\omega 2}) \cdot \sin\left(\frac{\pi}{2}\cdot(1-\tau_2)\right)\right) \quad (7)$$

$$\text{if } t \in [t_{ij3}, t_{ij4}] \text{.}$$

The notations τ_1 and τ_2 are:

$$\tau_1 = \frac{t-t_{ij1}}{t_{ij2}-t_{ij1}} \quad , \quad \tau_2 = \frac{t-t_{ij3}}{t_{ij4}-t_{ij3}} \quad (8)$$

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 appearence. 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.

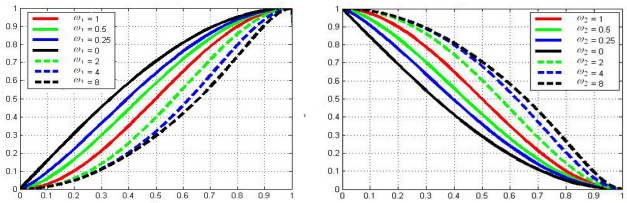

Figure 2: Various possible aspects for different ω_1 and ω_2 values

Figure 2 shows different possible aspects of the two parts of a viseme occurrence curve for a *j*-indexed appearance.

4. RESULTS AND CONCLUSIONS

We have applied our visual speech animation system on various test 3D virtual heads. Figure 3 shows one of the virtual actors we used for tests:

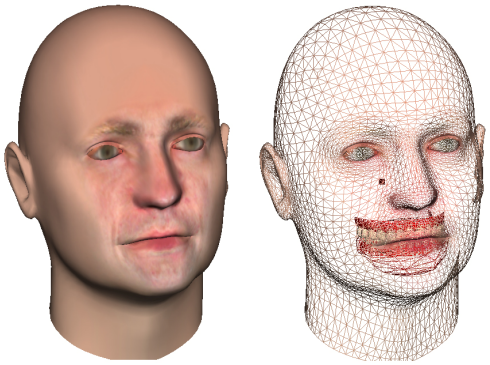

Figure 3: Example of a 3D virtual head used for tests in our work

For this particular head object, we used 25 visemes in order to thoroughly simulate visual speech production. Our animation model has proven its efficiency through several Romanian pronunciation tests, implying different coarticulation effects. For instance, Figures 4 and 5 present snapshots from the animation sequences of the Romanian words "*oală*" [o̯a.lə] and "*brom*" [brom] pronounced using our method and the 3D head from Figure 3.

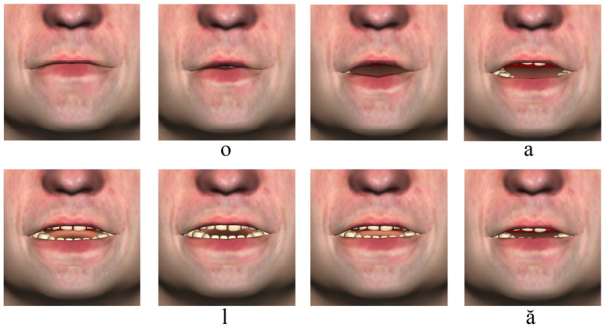

Figure 4: Visual speech shots of the virtual actor for the word *oală*

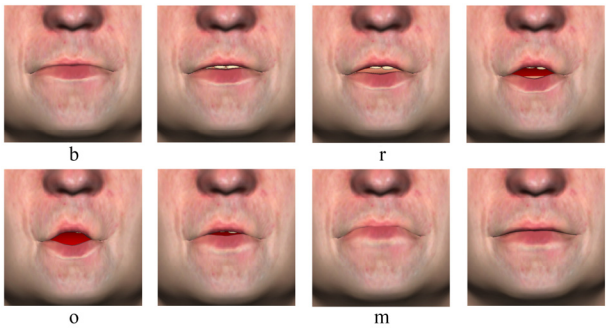

Figure 5: Visual speech shots while pronouncing the word *brom*

As seen in the two figures, the lip aspects while pronouncing *o* as a vowel in "*brom*" and as a semivowel in "*oală*" 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 consonant *r* is much affected by its following vowel, the consonant *l* needs tongue-to-teeth raising, and the labial consonants *b* and *m* imply total mouth closure. Figures 6 and 7 show plots of the viseme functions used for the two words from Figures 4 and 5.

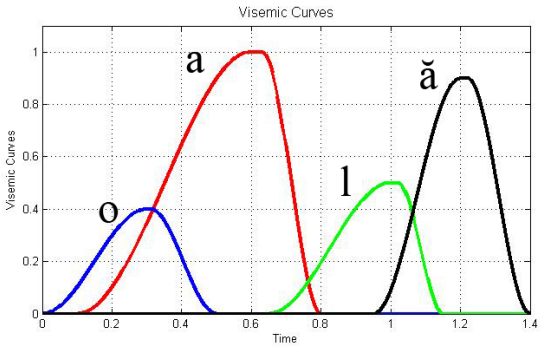

Figure 6: Viseme Dominance Curves for the word *oală* [o̯a.lə] from Figure 4

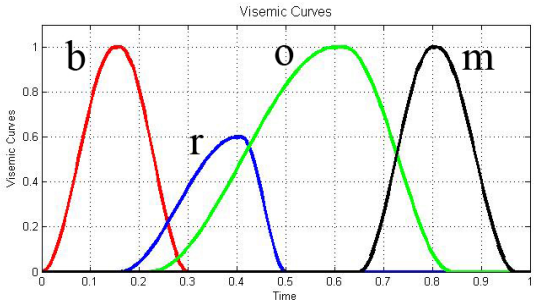

Figure 7: Viseme Dominance Curves for the word *brom* [brom] from Figure 5

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. Figure 8 shows tongue animation snapshots related to the virtual actor from Figure 3 while saying "*tractor*" [trakˈtor]. As seen in the picture, even if the lip shape for *t* differs when it is situated before *r* and *o*, the tongue position is quite similar for the two situations.

Figure 9 shows the corresponding front-view lip shapes for the two kinds of *t* from the word "*tractor*" in Figure 8.

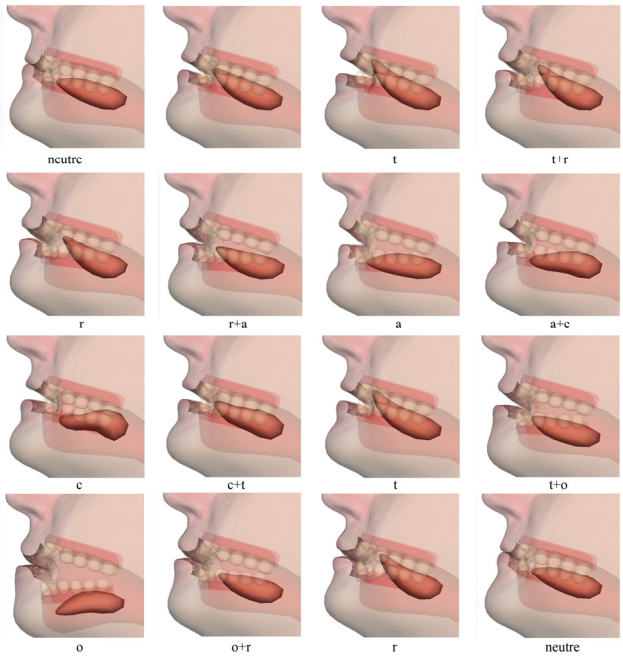

Figure 8: Tongue activity while pronouncing *tractor* [trak'tor]

Figure 9: The two *t* apparitions in the word *tractor* for Figure 8

The purpose of our work is to provide a framework designed for helping people suffering from severe or total deafening learn how to articulate correctly in the Romanian language. For this, our tongue movements observation feature is highly useful, allowing the user to closely study the entire oral cavity activity while different words are pronounced, such as in the example from Figure 8. Also, the speech playback speed is adjustable.

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.

Regarding future work, we intend to add more realism to the facial animation system by applying various changeable emotions to the virtual actors during speech, and also blinking effects. Our work could be very efficiently added to a Romanian language text-to-speech system, resulting into a Romanian-specific virtual talking head multimedia application.

ACKNOWLEDGEMENT

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

REFERENCES

[1] Frederic Parke. Computer generated animation of faces. In Proc. ACM Nat'l Conf., volume 1, pages 451–457, 1972.

[2] Frederic Parke, Keith Waters. Computer facial animation. AK Peters, 2008.

[3] C. G. Fisher. Confusions among visually perceived consonants. Journal of Speech and Hearing Research 1968, Issue 11, pages 796–804.

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

[5] Pushkar Joshi, Wen C. Tien, Mathieu Desbrun and Frédéric Pighin. Learning Controls for Blend Shape Based Realistic Facial Animation. In Proceedings of the SIGGRAPH '06, Art.17, 2006.

[6] Michael M. Cohen, Dominic W. Massaro. Modeling Coarticulation in Synthetic Visual Speech. In: Thalman N.M., Thalman D. (editors) Models and techniques in computer animation.Springer, Berlin Heidelberg New York, pages 139–156, 1993.

[7] Jyong Ma, Ronald Cole. Animating visible speech and facial expressions. In: The Visual Computer: International Journal of Computer Graphics, Volume 20, Issue 2, pages 86-105, Springer-Verlag, 2004.

[8] Fu-Chung Huang, Yu-Mei Chen, Tse-Hsien, Wang Bing-Yu, Chen Shuen-Huei Guan. Animating Lip-Sync Speech Faces by Dominated Animeme Models. In: Proceedings of the SIGGRAPH '09, Article no.2, New Orleans, 2009.

[9] Jose Mario De Martinoa, Leo Pini Magalhaaesa, Fabio Violaro. Facial animation based on context-dependent visemes. In: Journal of Computers & Graphics, pages 971-980, 2006.

[10] H. H. Bothe, F. Rieger. Visual speech and coarticulation effects. In: Proceedings of the IEEE ICASSP '93, pages 634-637, Mineapolis, USA, 1993.

[11] Alice Wang, Michael Emmi, Petros Faloutsos. Assembling an Expressive Facial Animation System. In: Proceedings of the 2007 ACM SIGGRAPH symposium on Video games, pages 21-26, 2007.

[12] Romanian Academy. DOOM - Dicționarul Ortografic, Ortoepic şi Morfologic al Limbii Române. Romanian Academy Printing House, second edition, 2010.