BULGARIAN LANGUAGE SPEECH AND LIP-SYNC ANIMATION

Vesselin Gueorguiev, Desislava Velcheva

Abstract: *The paper discusses a method for implementation of weight metamorphose and a lip-sync animation for speaking Bulgarian 3-D computer-generated realistic human head. There are explanation of experimental results with speaking head, basic dependences for Bulgarian language used in process of lip-sync animation and technology for implementation of lip-sync animation.*

Key words: *computer animation, lip-sync, realistic human head, 3D morphing.*

I. INTRODUCTION

At present the lip-sync animation is one of the most exploited kind of computer animations. All 3D computer-generated characters talk: realistic human heads, stylized heads, animals etc. This determines a need of defining the basic rules and methods for realization a head talking Bulgarian language.

The major problem for lip-sync animation is the visual plausibility. This is a result of millennial human experience. The communication among humans is the base of the civilization and every one has more experience how to understand face mimics of people around. The telephone is a too young kind of human communication and many people do not understand correctly speaking if they don't see the interlocutor. Papers [3] [4] show very indicative results illustrating that hearing is much successive if it is accompanied by video.

This paper is focused on illustration of a methodology and tests for lip-sync animation of realistic human heads with metamorphoses. For all models and animations we used 3DS MAX R5 without special plug-ins.

II. BULGARIAN LANGUAGE LIP-SYNC ANIMATION USING 3D MORPHING

A. Phonemes and Visemes

Before starting a creation a lip-sync animation by 3D weight morphing every computer animator should solve two major problems: that of phonemes and visemes [5].

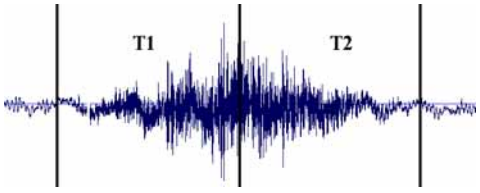
Every language has a limited number of phones which can be used for description of any kind of speech. These phones are named phonemes (the full definition is: "A phoneme is the smallest contrastive unit in the sound system of a language" [6]).

Every computer animator who starts to learn phonetics for his/her own needs to understand minimum three basic features for phonemes:

- ◊ A set of phonemes;
- ◊ Phonemes lengths for single phone, single word and in a sentence;
- ◊ Some special features of phonemes.

Complete explanation of these features of Bulgarian language is made in [1] [2]. The authors used "two-period" scheme for spoken phoneme and deduce dependences between phonemes while they are spoken in a single word or in a sentence. The model described by them uses:

- ◊ first period (T_1) for determining the influence between the current phoneme and the previous (glowed) phoneme;
- ◊ second period (T_2) for describing the influence between the current phoneme and the next one.



When the computer animation specialist is familiar with phonemes for a given language he/she should start to chouse the necessary visemes. The term "viseme" means the expression of a human face while he/she pronouncing single phoneme. The process of choosing visemes has specific features, some of which are directly relevant to visual plausibility of lip-sync animation. These specific features are:

- ◊ Lips configuration: for everyone this is specific but statistically people who speak one and the same language display a similar lips configuration for a phoneme;
- ◊ Visible organs of speech: how these organs take part in the process of phoneme pronounce;
- ◊ Facial muscles and phonemes: grouping phonemes by means of muscles that are used for pronunciation;
- ◊ A place of atriculation for all phonemes: the place of articulation is a zone with maximal contraction of vocal duct when people pronounce a phoneme (the tongue participates in the process of phone construction);
- ◊ Phoneme articulation type for a consonant between two vowels.

The phonemes in Bulgarian language can be grouped basically in three groups if the stated above specific features are used:

- ◊ Phonemes which are pronounced with similar lips configuration and similar tongue position. The difference between these phonemes is only in vocal chords usage during the process of phone generation (voiced and voiceless consonants). Example: Bulgarian phonemes *[б]* and *[п]*.
- ◊ Phonemes engaging vocal organs in one and the same way except tongue position.
- ◊ Phonemes which are pronounced with great differences on the face expression.

This grouping of phonemes allows us to generate a set of 12 visemes for Bulgarian language lip-sync animation:


[А]


[Е]


[И]


[О]


[У]


[Ф]


[П]


[С]


[Ш]


[Та]


[То]


[Л]

Our tests with these visemes demonstrated that this set is suitable for big area of lip-sync animations.

B. Visemes and 3D weighed morphing for Bulgarian language: analysis and implementation rules

The experiments with based on lip-sync animation Bulgarian speaking head led to the following results:

1. The variation of viseme's influence between keyframes should be realized by a curve described by second or third order polynomial. This realization has two basic advantages:
   - ◊ Changes in human's face in speaking process are fluent. This is very important for visual plausibility.
   - ◊ The use of curve's tension and bias speech dynamics can be controlled. This is important when should be realized human's emotion. Combining "slow-in/slow-out" techniques human's individuality can be displayed.
2. Visemes influences' curves can be overlapped in process of speech generation. The reason is that in the moment when people pronounce a phone they are preparing themselves for pronunciation of the next phone. This enables image designer to generate persons having individuality. The classical overlap is: when the viseme has maximum influence (100%) a previous viseme's influence reaches to zero. We recommend that a variation up to 7% for the fading viseme should be applied.
3. The tongue has great influence upon the face of speaking people. The most suitable viseme for visualizing this influence is [Л].
4. There are no visemes for [к, г, х] phonemes because when people pronounce them lips' and tongue's visual configuration are not changed
5. The vowels' influence starts from the beginning of antecedent consonant and comes to an end in the maximum of the next phonemes. The reason for this is consonant's liability.
6. When the phonemes [л] and [р] are between vowels, the vowel's transitions are overlapped using the basic rule.
7. Visemes ([П], [Ф], [Ш]) have very strong alterative effect upon the face and this can be realized when using this visemes. Our recommendation is to increase their length by 3 periods before and 3 periods after the maximum.
8. The viseme [С] has less alterative effect than the visemes ([П], [Ф], [Ш]) but this alteration is greater than that of the standard visemes. Our recommendation is to increase its length by 2 periods prior and 2 periods after the maximum.
9. When a viseme has to be used for visualization consecutively its minimal influence should not be 0% at the time of the maximum influence of the middle viseme. This influence is specifically for persons for different visemes but statistically it is 3-8%. This nonzero influence increases visual plausibility because makes speaking more fluent.
10. The influence of the first viseme after a pause should begin 4-5 frames (0.085-0.108s) prior to reaching its maximum influence. The reason for this is the process of mouth opening before the beginning of speech. The additive time is specifically for persons but for a single person variation between phones is 5-8%. The emotion's influence is greater (up to 35%).
11. The length of the last viseme prior to a pause should be 2 or 3 (0.062 - 0.08 s) frames greater. The reason for this is the process of mouth closing: when the mouth is closing the sound for the last phone extends. The additive time is specifically for persons but for a single person variation between phones is little (3-5%). The emotion's influence is greater (up to 40%).

C. Bulgarian language lip-sync animation: generation process

We propose a sequence of steps for successful generation of lip-sync animation for Bulgarian language. Our generation process is based on test results described above:

1. Creation a 3D human's head in an "at rest" pose. This is a "basic" model.
2. Generation of all visemes' targets using the "basic" model. In weighed 3D metamorphose the "basic" model and visemes' targets get combined in the generation process. The viseme target's influences (represented in %) specifies how much the selected intermediary viseme target contributes to the overall morph solution.
3. Measurement of the length of the basic phonemes [и, а, о, п, с, ш, р]: the voice of the person who will be used for animated character is used (example: Table 1).

Table 1.

phonemes	T1 [sec]	T2 [sec]
И	0,085	0,085
А	0,075	0,075
О	0,081	0,081
П	0,046	0,046
С	0,135	0,135
Ш	0,082	0,082
Р	0,055	0,055

4. Generation a table of lengths of all language phonemes using dependency between basic phonemes and other phonemes of the language (Table 2).

Table 2.

phonemes	T1 [sec]	T2 [sec]	Length
И	0,085	0,085	0,17
Е	0,092	0,092	0,184
А	0,075	0,075	0,15
Ъ	0,055	0,055	0,11
О	0,081	0,081	0,162
У	0,076	0,076	0,152
П	0,046	0,046	0,092
Б	0,046	0,046	0,092
П', Б'	0,046	0,097	0,143
Ф, В	0,092	0,092	0,184
Ф', В'	0,092	0,194	0,286
М	0,056	0,056	0,112
М'	0,056	0,056	0,112
С	0,135	0,135	0,27
Т, Д пред А, Ъ	0,059	0,043	0,102
Т, Д пред О, У	0,059	0,043	0,102
Т, Д пред И, Е	0,059	0,059	0,118

Т' пред А, Ъ	0,059	0,178	0,237
Т' пред О, У	0,059	0,178	0,237
Т' пред И, Е	0,059	0,227	0,286
Д' пред А, Ъ, О, У	0,059	0,163	0,222
Д' пред И, Е	0,059	0,21	0,269
З	0,135	0,135	0,27
С', З '	0,135	0,189	0,324
Ц, ДЗ	0,081	0,189	0,27
Ц'	0,081	0,173	0,254
Л	0,059	0,059	0,118
Л'	0,059	0,167	0,226
Ш	0,082	0,082	0,164
Ж	0,082	0,082	0,164
Ч, ДЖ	0,093	0,093	0,186
К, Г	0,077	0,077	0,154
К', Г'	0,077	0,207	0,284
Х	0,148	0,148	0,296
Х'	0,148	0,278	0,426
Р	0,055	0,055	0,11
Н	0,055	0,055	0,11
Н', Р'	0,055	0,055	0,11

5. Personnel peculiarities of speech must be analyzed:
   - ◊ The way this person combines phonemes in a single syllable and syllables in a word;
   - ◊ What is his normal tempo of speaking;
   - ◊ The way emotion changes speaking tempo.
6. Determination of changes in vowels' length: this analyzes displays the dependences between vowels and their positions in a single word; between vowel's position and the point of the accent. The result of these tests is a table which contains vowels' length reduction (example: Table 3).

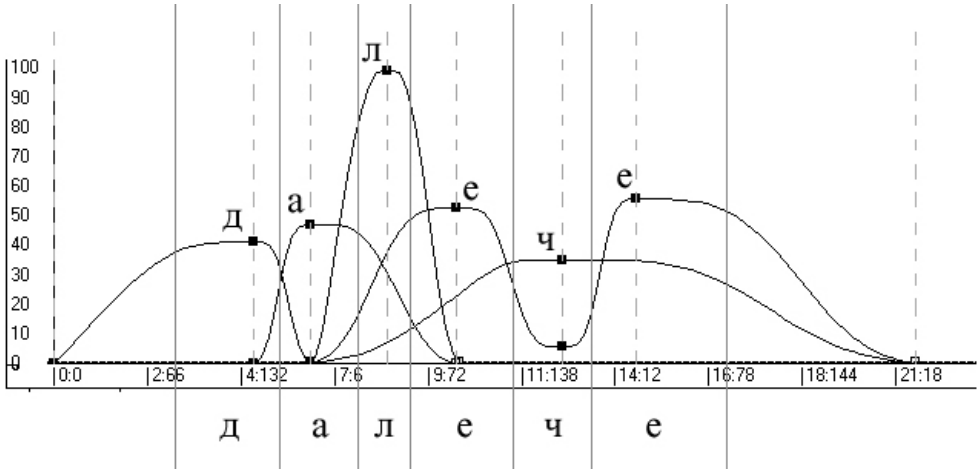
Table 3.

vowels	In 1st accented syllable	In 2nd accented syllable	In 3rd accented syllable	In 1st syllable before accent	In 2nd syllable before accent	In syllable after accent	In last syllable
а	+ 5%	+ 21%	+ 10%	- 2%	- 8%	- 8%	+ 5%
ъ	+ 4%	+ 18%	+ 10%	- 2%	- 8%	- 8%	+ 5%
о	+ 3%	+ 15%	+ 8%	0%	- 5%	- 7%	+ 3%
у	+ 2%	+ 14%	+ 8%	0%	- 5%	- 6%	+ 2%
е	+ 2%	+ 3%	0%	0%	-2%	0%	+ 5%
и	0%	0%	0%	0%	- 1%	0%	+ 2%

7. Determining visemes' influence during of speech generation. This influence is based upon phonemes' lengths. Our tests indicate that a good practice for Bulgarian language is :
   - ◊ Vowels: the base length is phoneme's length in an unaccented syllable;
   - ◊ Consonants: the base length is the average length of a phoneme spoken separately.

   When a vowel's and consonant's length is determined viseme's influence is calculated. The influence depends on viseme's type: opened or closed. After

calculation minimal and maximal viseme's influence for current phonemes generate a table of viseme influence in different combination. This defines the degree of opening the mouth while pronouncing the current phone. *Example*: The way the word „далече" ("far away") is pronounced in a sentence. In test the word is the first word in the sentence and emotion is lyrical.



8. Correcting visemes' length using words length from audio channel. This correction will be proportional.
9. Viseme's length and influence is used for setting 3D morphing parameters.

III. CONCLUSIONS AND FUTURE WORK

This paper presents results of research oriented to generation of speaking Bulgarian language realistic computer-generated head based on 3D morphing. Experimental results demonstrate that this model is suitable for this application but needs additional research for emotion visualizations.

IV. REFERENCES

[1] Boyadzhiev T, Tilkov D., "*Literature Bulgarian Language Phonetics"* (Бояджиев Т., Д. Тилков, *"Фонетика на българския книжовен език"*), 1999.

[2] Stoyanov, S. *"Literature Bulgarian Language Grammar"* (available only in Bulgarian : Стоянов, Ст., *"Граматика на книжовния български език"),* 1984.

[3] Cohen M. ,Massaro D., "*Modeling coarticulation in synthetic visual speech",* 1993

[4] Breeuwer, M., Plomp, R., "*Speechreading supplemented with formant- frequency information for voiced speech",* 1985

[5] M. Comet, "*Lip Sync – Making Characters Speak"*, 1998

[6] http://www.sil.org/linguistics/GlossaryOfLinguisticTerms/

V. ABOUT THE AUTHOR

Ass.Prof. Vesselin Gueorguiev, MagEng., Department of Computer and Control Systems, Techinical University of Sofia, Phone: +359 2 9652192, E-mail: veg@tu-sofia.bg

Ass.Prof. Arch. Desislava Velcheva, Department of Informatics, New Bulgarian University, Phone: +359 2 8110658, E-mail: dvelcheva@nbu.bg