A Model of Personalized Audio-Visual TTS-synthesis for Russian

B. Lobanov,* L. Tsirulnik,* A. Ronzhin,** and A. Karpov**

*United Institute of Informatics Problems of NAS Belarus
**St. Petersburg Institute for Informatics and Automation of the RAS
{lobanov, l.tsirulnik}@newman.bas-net.by
{ronzhin, karpov}@iias.spb.su

ABSTRACT

The paper deals with the description of peculiarities of audio-visual TTS-synthesis model for Russian. The developed viseme classes are given, the method of audio- and visual flows synchronization is outlined. The audio-visual TTS-synthesis system, built on the basis of the model described, can be used as a multi-speaker and multi-language system.

1. Introduction

The tendency of developing speech technologies points out the relevance of including visual information as an additional channel of speech perception and recognition [1]. Visual information is very important for speech recognition in noisy environments and is indispensable for people with restricted hearing or defects of pronunciation. The number of research in the fields of audio-visual speech recognition and TTS-synthesis is permanently increasing.

There are two approaches to the creation of audio-visual TTS-synthesis systems (often called „talking head"): imitation [2, 3] and concatenative [4, 5]. In the imitation approach a 2D or 3D model of the head and face is created and parameters for facial expression and lips movement representation are adjusted. In the concatenative approach the "talking head" is generated by choosing corresponding video fragments or images from the visual database (DB) of a certain speaker.

The advantage of the imitation approach is a smaller physical size of the data necessary for visual speech synthesis. On the other hand, the implementation of a 2D or 3D model presents considerable computational complexity. Moreover, the imitation approach does not give sufficiently realistic results in the personification of the "talking head" due to unavoidable sketchiness of speech movement representation. Thus, the concatenative approach is more preferable for the purposes of creating a system of personalised audio-visual TTS-synthesis.

The important tasks arising in the creation of an audio-visual TTS-synthesis system are the development of visual units of speech – visemes – and synchronization of audio and visual flows. Until the present time systematic research in these directions for Russian has not been carried out.

The audio-visual personalized TTS-synthesis model, described in the present paper, is the result of further development of the theory and technology of personal voice cloning [6]. The paper is organized as follows: in part 2 the creation of visemes for Russian is described; part 3 contains the description of audio and visual flows synchronization; the method of synthesising audio-, as well as visual speech segments transition points, is outlined in part 4.

2. Visemes Creation for Russian Speech

The determination of the minimally sufficient set of visemes is based on the known classification of Russian phonemes according to the manner and place of articulation subject to co-articulation and reduction effect [7]. As the research shows, for Russian speech the tongue body, tip and sides, as well as the uvula and vocal cords dynamics of movements are completely hidden. Only the lips and lower jaw movements are visually accessible. These movements are manifested most clearly in the formation of vowels (fig. 1), as well as labial consonants (fig. 2).

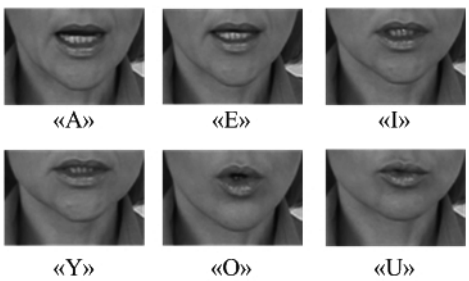

Figure 1. Visemes of vowel phonemes.

Figure 2. Visemes of groups of consonant phonemes.

Less significant, but noticeable enought difference is observed between hard and soft consonants, as well as between velar and blade consonants (see fig. 2).

Based on described phenomena, a minimally sufficient set of Russian visemes is selected (Table 1). The indices of vowels point to the degree of the vowels' positional

Table 1. "Phoneme-viseme" correspondence for audio-visual Russian speech synthesis

Viseme	Allophones of phonemes	Viseme	Allophones of phonemes
V_0	# (pause)	V_8	B', P', M'
V_1	A_0, A_1	V_9	F, V
V_2	E_0, E_1	V_{10}	F', V'
V_3	I_0, I_1, I_2, I_3	V_{11}	C, S, Z, Sh, Zh, D, T, L, R, N
V_4	O_0, O_1	V_{12}	C', S', Z', Sh', D', T', L', R', N'
V_5	U_0, U_1, U_2, U_3	V_{13}	G, K, H
V_6	Y_0, Y_1, Y_2, Y_3, A_2, A_3, E_2, E_3	V_{14}	G', K', H', J'
V_7	B, P, M		

reduction: 0 – fully stressed vowel, 1 – partially stressed vowel, 2 – prestressed vowel, 3 – poststressed vowel. The symbol "'" after the consonant indicates its softness.

The process of viseme images creation can be carried out in two ways: by using video fragments of a speaker pronouncing a phonetically balanced text corpus (for example, the text from the appendix 1 in [7]), or by creating snapshots of a face of a speaker imitating the pronunciation of one or another sound corresponding to this or that viseme. As the experience of visemes creation shows, the second way is preferable because of smaller laboriousness of visemes creation, as well as the possibility of creating a more reliable fixation of standard head position.

3. Visemes and Viseme Transitions Duration Settings

The synchronization of viseme display with the synthesized speech signal is realized on the basis of information about the position of the beginning and the end of each allophone in the current speech flow. On the visual level it is nessesary to specify three sections, whose total duration is equal to the real duration of the current allophone t_a: the initial transition t_i, the stationary section t_s and the final transition t_f.

The dynamic process of stationary and transition sections displaying one of the viseme parameters – the degree of lip opening – is shown in fig. 3. The dynamic visual display of *i*-th allophone production is formed by displaying the initial, stationary and final sections image sequence. Hereby the total duration of a transition from (*i*–1)-th allophone to *i*–th allophone t_t is formed by the final and initial transition sections.

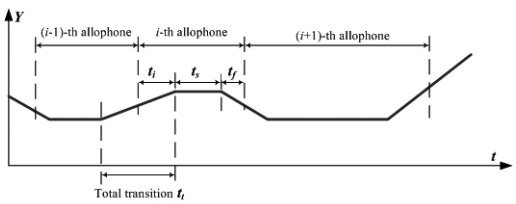

Figure 3. The process of viseme sequence display (*Y* is the degree of lip opening, *t* is the time).

Thus, the task is to specify the current values of t_i and t_f, estimated by the number of frames. The duration of each frame display is equal to 40 ms (taking into account that the standard rate of frames refresh is 25 frames per second).

The current duration of every allophone ta is set by the audio TTS-synthesis system on the basis of an allophone average proper length and required speech tempo. Table 2 demonstrates the values of allophones relative duration (in %) when changing the speech rate, as well as the allophones absolute duration in milliseconds (ms) and number of frames (frm).

Table 2. The relative (%) and absolute (ms, frm) durations of phonemes in different speech rate

No	Speech items type	Slow rate (% – ms – frm)	Medium rate (% – ms – frm)	Fast rate (% – ms – frm)
1.	Pause	250 – 650 – 16	100 – 260 – 7	20 – 50 – 2
2.	Stressed vowels	200 – 320 – 8	100 – 160 – 4	50 – 80 – 2
3.	Prestressed vowels	200 – 160 – 4	100 – 80 – 2	80 – 64 – 2
4.	Poststressed vowels	200 – 80 – 2	100 – 40 – 1	80 – 20 – 1
5.	Sonants	140 – 110 – 3	100 – 80 – 2	80 – 64 – 2
6.	Voiced plosive and fricative consonants	120 – 120 – 3	100 – 100 – 3	80 – 80 – 2
7.	Unvoiced plosive consonants	130 – 160 – 4	100 – 120 – 3	85 – 100 – 3
8.	Unvoiced fricative consonants	130 – 180 – 4	100 – 140 – 4	85 – 120 – 3

As can be seen from table 2, the shortest allophones can be displayed by only one frame. This frame should correspond to the stationary section; the initial and final transitions are not present. That does not mean, hovewer, that durations of the initial and final transitions are equal to zero for (i–1)-th and (i+1)-th allophones, thus, as a rule, the transition process can be displayed even in the fast rate of synthesized speech.

When calculating the stationary sections of visemes it is nessesary to take into account not only the information about the start and end positions of the allophone and its duration t_a, but also the well known phenomenon of vowels and consonants coarticulation [7]. On the visual level this phenomenon reveals itself in the fact that in „consonant-vowel" syllables the vowel-specific articulational movements are set not only on the vowel, but also on the most part of the consonant phoneme. The expression of co-articulation phenomena is different for different „consonant-vowel" combinations. For the velar consonants – /H/, /G/, /K/ – co-articulation is expressed in combination with any vowel, while for other consonants this phenomenon is expressed only in combination with labial vowels – /U/ and /O/. Thus, the graphical representation of sequence of frames of visemes and transition sections, shown in fig. 3, only illustrates the cases when the co-articulation effect is not observed (in words, where labial vowels and velar consonants are absent). Fig. 4 demonstrates the appearence of co-articulation phenomena by the example of the Russian phonetic word *в кассу* (English translation – *to the cash desk*, phonemic transcription /FKASSU/). This word contains the combination of the velar consonant /K/ – the vowel /A/ and the consonant /S/ – the labial vowel /U/. The arrows in fig. 4 point to the stationary segments of corresponding visemes, the slanting

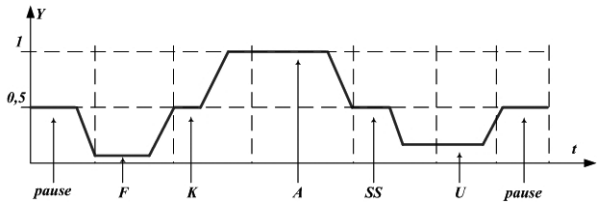

Figure 4. The illustration of co-articulation processes.

lines show to the transition segments, the dashed lines indicate the start and end points of synthesized phonemes.

After calculating the duration and positioning the phoneme stationary section subject to co-articulation it is nessesary to specify the values of transition durations t_i and t_f. As the experience shows, satisfactory results at medium speech rate can be reached by settings of ti and tf values equal to 70–90 ms (2 frames). These values can be reduced or increased depending on the required speech rate.

4. Synthesis of Transition Sections of Audio- and Visual Speech Segments

For the reconstruction of the continuous movement of articulatory organs by static images the computer-generated imagery methods are used in [5]. In the present work this purpose is achieved by another significantly simpler procedure, the so called Soft Lacing (SL) procedure, which provided good results in sound waves lacing in the microwave TTS-synthesis system [8].

The sound waves lacing is used to create smooth sound transition between two speech segments. For the smooth transition one wave period of the first and second signals are used. The transition section generation is performed by continuous damping the first signal and amplification of the second signal, as it is shown in fig. 5.

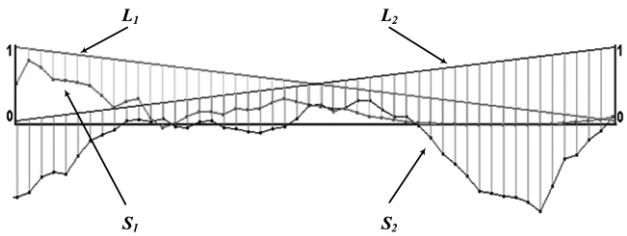

Figure 5. Generation of SL-transition sections.

Damping of the first signal is calculated using the formula:

$$S'_1(n) = S_1(n)L_1(n), \quad \text{where } L_1(n) = 1 - \frac{n}{N}, \quad 0 \le n \le N \tag{1}$$

where $S'_1(n)$ is the modified signal 1, $S_1(n)$ is the initial signal 1, N denotes the size of smooth transition frame.

Amplification of the second signal is expressed by the formula:

$$S'_2(n) = S_2(n)L_2(n), \quad \text{where } L_2(n) = \frac{n}{N}, \quad 0 \le n \le N \tag{2}$$

where $S'_2(n)$ is the modified signal 2, $S_2(n)$ is the initial signal 2.

For the calculation of the summarized signal of transition section $S'(n)$ the following formula is used:

$$S'(n) = S'_1(n) + S'_2(n), \quad 0 \le n \le N \tag{3}$$

The described procedure of period-by-period sound signals soft lacing corresponds to the frame-by-frame soft lacing of face visual frames while pronuncing the corresponding sounds.

In computer graphics the process of fusion of two images, in order to create the partial transparency effect, is known as weighted superposition of blended colors or alpha-blending [9]. Alpha-blending is a pixel-by-pixel blending of source and background color data. Each of the three components (red, green, blue) of a given source color is blended with the corresponding component of the background color according to the following formula (taking into account that each color component is presented by one byte):

$$C = \frac{\alpha C_1}{255} + \frac{(255 - \alpha)C_2}{255}, \quad 0 \le \alpha \le 255 \tag{4}$$

where α is the weight of the image C_1 relatively to the background image C_2, The alpha value indicates the transparency of the color — the extent to which the color is blended with the background color.

For the viseme transition sections the values of α for each of the RGB-components are equal and calculated according to the formula:

$$\alpha(n) = 255 - \left[255\frac{n}{N}\right], \quad 0 \le n \le N \tag{5}$$

where n is the number of the current frame of a transition section, N denotes the total number of frames of a transition section.

5. Conclusion

The personalized “talking head” model for Russian is presented in the paper. The software implementation of the model for Windows-family operating systems is created and used as a component of the audio-visual TTS-synthesis system [10]. The peculiarity of system implementation is simultaneous work of the modules of visual and acoustic data processing. Synchronization of audio- and visual flows on the software-based level is implemented by the standard MS Windows means for multithreaded processing.

The speaker-dependent and language-dependent data and the conversion rules of the audio-visual TTS-synthesis system are stored in special DBs, that allows the system to be used as a multi-speaker and multi-language one provided that the corresponding linguistic, acoustical and visual resources are supplied.

Audio-visual Russian synthetic speech provided by the system will be demonstrated at the conference.

The research is supported by Russian and Belarussian Foundations for Fundamental Research under the grants No 08-07-90002 –No Ф08Р-016.

BIBLIOGRAPHY

[1] *Issues in Visual and Audio-Visual Speech Processing* (2004), Cambridge: MIT Press.

[2] Tekalp A. M., Ostermann J., (2000) *Face and 2-D mesh animation in MPEG-4*, in: "Signal Processing: Image Communication, Special Issue on MPEG-4. Volume 15".

[3] Chen L.S. et al., (1997) *Animated talking head with personalized 3D head model*, in: "IEEE First Workshop on Multimedia Signal Processing".

[4] Bregler. C et al., (1997) *Video Rewrite: Driving visual speech with audio*, in: "Proc. DIGGRAPH97".

[5] Cosatto E., Graf H.P., (2000) *Photo-realistic talking-heads from image samples*, in: "IEEE Transactions on Multimedia. Volume 2".

[6] Lobanov B., Karnevskaya H., (2002) *TTS-Synthesizer as a Computer Means for Personal Voice Cloning (On the example of Russian)*, in: "Phonetics and its Applications. Stuttgart: Steiner".

[7] Lobanov B., Tsirulnik L., (2008) *Computer synthesis and cloning of speech*, Minsk: Belorusskaya Nauka, 334 p. (in Russian).

[8] Lobanov B., (1991) *MW-Speech Synthesis from Text*, in: "Proc. XII International Congress of Phonetic Sciences ICPhS'91", Aix-en-Provense.

[9] Porter Th., Duff T., (1984) *Compositing Digital Images*, in: "Computer Graphics. Volume 18 (3)".

[10] Karpov A., Lobanov B., Ronzhin A., Tsirulnik L., (2008) *Audio-Visual Russian Speech Recognition and Synthesis for a Multimodal Information Kiosk*, in: "The Fifth International Conference on Neural Networks and Artificial Intelligence", Minsk.