ISSN 1392 – 124X INFORMATION TECHNOLOGY AND CONTROL, 2009, Vol. 38, No. 3

ENGLISH TALKING HEAD ADAPTATION FOR LITHUANIAN SPEECH ANIMATION

Ingrida Mažonavičiūtė, Romualdas Baušys

Vilnius Gediminas Technical University
Sauletekio al. 11, LT-10223 Vilnius, Lithuania
e-mail: gsk@fm.vgtu.lt

Abstract. Visual speech animation plays an important role in human-computer interaction. To force already existing English Talking Head speech animation engine to talk Lithuanian, some modifications to the animation script were made. For this adaptation, the relation between English and Lithuanian languages was explored. To determine it, 30 3-dimensional Lithuanian visemes were modeled using the calibration of two orthogonal pictures of phoneme. Using the visual similarity of different English and Lithuanian phonemes, "Lithuanian phoneme to English viseme mapping table" was defined and used for Lithuanian speech animation.

Keywords: talking head, speech animation, viseme, phoneme, phoneme to viseme mapping table.

1. Introduction

Human communicate using words and sentences; visual information, such as facial expressions, lips and tongue movements improve the perception of the uttered audio signal. Impaired hearing individuals can achieve very good speech perception because of lip-reading, but all people use speech reading. This is especially true when the acoustic conditions are inadequate. It has been showed that addition of the visual information increases the intelligibility with 57% for consonants, 30% for vowels, 39% for monosyllabic words, and 17% for short phrases [16].

The design and implementation of three dimensional (3D) synthetic head models that can produce naturally looking audio to video mapping („talking head") is still one of the challenging objectives of Human – Computer Interaction research. Despite the importance of synthetic speech animation in movie, advertising and computer game industries, "Talking head" can be widely used for interactive applications, where User Interface agents can be developed to be employed in e-learning, Web navigation or as virtual secretary [14]. But the most important thing is that hearing-impaired people can benefit from synthetically generated talking faces by means of visual speech, for instance videophones can be produced to make possible the distant communication of the deaf people [9].

In this paper the adaptation of the English speech recognition system for the Lithuanian speech animation was presented. Many authors of "talking head" software claim that their models are speech independent, but practically audio driven facial animation requires training of a speech recognition system which is used for generating phoneme and viseme alignments from the input speech. For this reason we have to explore both visual and acoustical aspects of Lithuanian speech and to construct the mapping table between the Lithuanian phonemes and the English visemes. The table will be used for Lithuanian speech animation.

This paper is organized as follows. In Section 2, we describe the visual speech animation generation methods, restricting our attention to viseme driven approach and its Audio to Visual mapping levels. In Section 3, translingual speech synthesis is presented. There we describe, in detail, the method for the adaptation the speech recognition system of one language to generate phonetic and visemic alignments in a new language. The specific case of adding Lithuanian words to an English speech recognition system is considered. In Section 4, the experiment and our mapping table for transmission of Lithuanian phonemes to English visemes is described. The mentioned table is presented here too. Finally, conclusions are presented in Section 5.

2. Background

For the generation of naturally speaking "talking head", positions of the mouth and tongue must be related to characteristics of the speech signal. Visual speech animation generation can be classified into two

different categories: data-driven approaches and viseme-driven approaches [5].

Data-driven approaches generate speech animations by concatenating pre-recorded facial motion data or sampling from statistical models learned from the data. These approaches typically produce realistic speech animation results, but it is hard to predict how much motion data are enough to train statistical models or construct a balanced facial motion database. In other words, the connection from the amount of pre-recorded facial motion data to the realism of synthesized speech animations is not clear. Furthermore, these approaches often do not provide intuitive process controls for the animator.

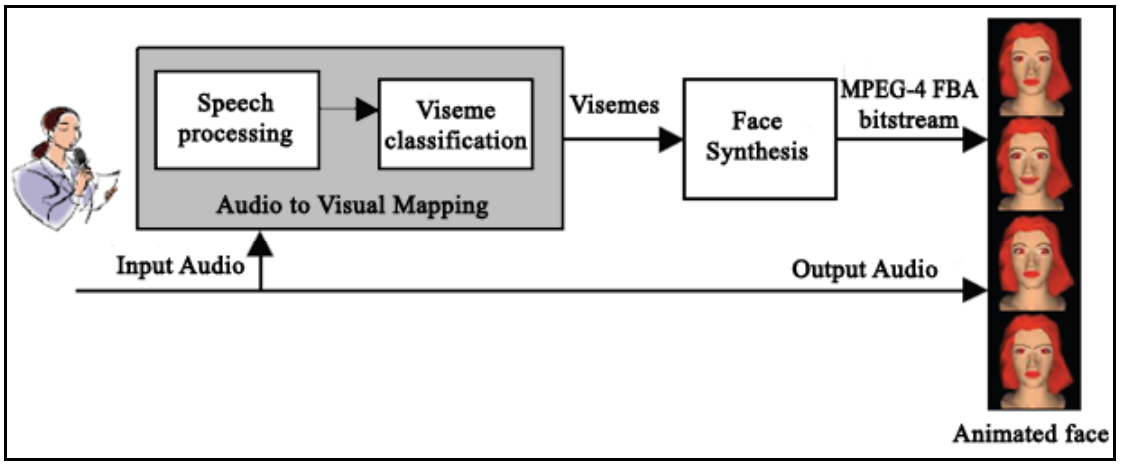


Figure 1. Schematic view of lip sync system

Viseme-driven approach is based on the fact that there are many sounds that are visually ambiguous when pronounced. Therefore, there is a many-to-one mapping between phonemes and viseme (basic visual unit that corresponds to the phoneme in speech).

Generation of novel speech animations using viseme-driven approach consists of 2 main steps [21]: first of all, visemes must be designed. Later visemes are combined using empirical smooth functions [4, 11] or co-articulation rules [1, 2].

The base idea of viseme-driven approach is demonstrated in Figure 1. As we can see, audio to visual transmission (mapping) is a key issue in bimodal speech processing due to the fact that the correctness of calculated visemes significantly influence the realism of speech animation. The important Audio to Visual (AV) mapping can be solved on several different levels [8]:

- signal level (front end),
- phoneme level (acoustic model),
- word level (language model).

Signal level concentrates on a physical relationship between the shape of the vocal tract and the sound that is produced. Speech signal is segmented into frames. The mapping is performed from acoustic to visual feature frame by frame. There are many algorithms that can be modified to perform such mapping – Vector Quantization (VQ) [21], the Neural Network (NN) [13], the Hidden Markov Model (HMM) [3].

Using phoneme level, speech is first segmented into a sequence of phonemes. For each phoneme, AV mapping is generated using a lookup table, which contains one visual feature set for each phoneme. The standard English set of visemes is specified in MPEG-4 and usually contains 15 static visemes that can be easily distinguished [17].

The language model is more concerned about context cues in the speech signals. First of all, speech is segmented into words, and then a HMM can be created to represent the acoustic state transition in the word. The first signal level techniques are incorporated in AV mapping for each higher hierarchical state.

The choice of the particular level depends on application where AV will be used. Signal level is simple, language independent and suitable for real-time implementation, but contrary to the latter two approaches, co-articulations are not incorporated. Both of phoneme and word levels are providing more precise speech analysis and depend from context. Due to the fact that higher input signal requires more complex speech recognition system, phoneme level is faster and simpler then word level. But there is one disadvantage: different phonemes are defined in different languages, so there is no one standard phoneme set, and speech animation engine must be revised for every new language.

3. Translingual visual speech synthesis

The speech recognition and animation engine is a critical part of any speech animation system. Building a speech recognition system is data intensive and is a very tedious and time-consuming task [6]. So it is very important to explore the possibility to use the speech animation engine of the *base language* (language used in training the speech recognition system) to animate the new language in which the video has to be synthesized (*novel language).* In this paper Lithuanian is the novel language and English is the base language.

If the input audio is in the same language as the language used to train the recognition system and audio file with its transcription were used, phonetic alignment is fine. But if the novel language word is presented to the speech recognition system which is trained in the base language, alignment fails to give the phonetic base forms of the word. This situation arises due to the fact that the base language vocabulary does not include words from the novel language. Since the recognition system is trained on the phone set of the base language, the vocabulary needs to be modified so that the words from the novel language would represent the base forms using base language phone set.

Because the aim of mapping the phone set is to generate the best phoneme boundaries through acoustic alignment, the mapping is based on acoustically-similar phonemes, i.e., if there is no phoneme in the base language which can be associated with the phoneme in the novel language, then that base language phoneme is chosen which is acoustically closest. Both, however, may map to a different viseme. So, mapping based on acoustically similar phonemes may distort the visemic alignment, as it does not take into consideration the visemes corresponding to each such phoneme.

Since the system has to work for any novel language using the alignment generator and the viseme set in the base language, visemic alignment cannot be simply generated from the phonetic alignment using direct phoneme to viseme mapping. An additional vocabulary based on the visual similarity of the two phonemes (in base and novel languages) has to be created. This mapping based on visemic similarity is called the visemic vocabulary modification layer. Using this additional vocabulary, the base language alignments and the base language phoneme-to-viseme mapping, we get the visemic alignments, which are used to generate the animated video sequence. Alternately, if the viseme set images are available for the novel language, then the visemic vocabulary modification layer can be modified to directly give the visemic alignment using the phoneme-to-viseme mapping in the novel language. If the viseme set of the novel language is very different from the viseme set of the base language, then this modified system would be especially useful. So the goal of this paper is to create English-Lithuanian visemic vocabulary table.

4. English - Lithuanian visemic vocabulary creation

The purpose of the paper is the adaptation of the English recognition and animation engine for the Lithuanian speech animation. For this task, we have used the English speech recognition and animation engine "Crazy Talk", which is one of the best commercial software in this area [20]. Also it was chosen due to the fact that the engine detects the phoneme time positions of Lithuanian speech file enough precisely (associated visemes were chosen mistakenly).

English phoneme set consist of 48 phonemes (this count varies). Standard Lithuanian alphabet consists of 32 characters, but there is different count of phonemes. According to Lithuanian grammar rules Lithuanian phoneme set consists of 58 units. We related Lithuanian and English phonemes according to the table proposed by Kasparaitis [10].

We have analyzed 30 Lithuanian phonemes transcribed by SAMPA standard [19]. Many acoustic sounds of separate languages are visually ambiguous and accordingly different phonemes can be classed using the same viseme. There is therefore a many-to-one mapping between Lithuanian phonemes and English visemes, so with a small bias, Lithuanian phonemes can be grouped into 16 different visemes defined in "Crazy Talk" (Figure 2).

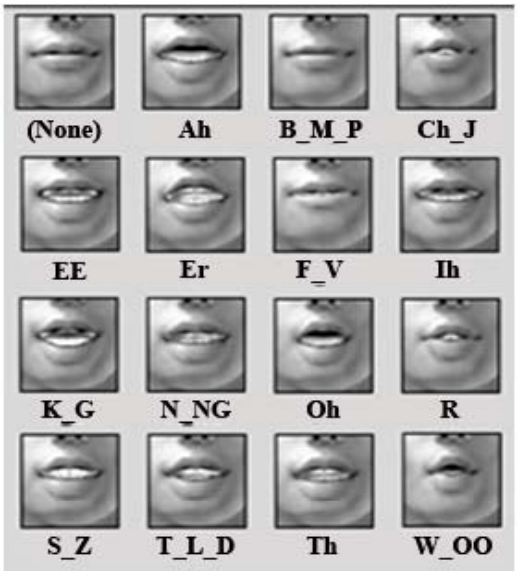


Figure 2. Visemes classification in „Crazy Talk" software

As we can see, visemes for such Lithuanian letters as ą,č,ę,ė,į,š,ų,ū,ž are not defined at all, so the animation can be unpredictable. But as we mentioned above, timing information of phonemes is quite good, so if we would know what English viseme must be placed at the detected time of particular Lithuanian sound, we could easily replace the wrong English viseme with the right one, selected for Lithuanian speech animation. Therefore we need to create additional English-Lithuanian visemic vocabulary.

4.1. Data collection

When we are exploring the relation between speech visual and acoustical aspects, it's very important to get the information of actual motion of the points in the face. There are two types of markers that can be used when recording facial movements: area markers and point markers [15]. The databases used in this study were recorded using point markers which were drawn on the face of the subject. The points were

chosen at the positions defined by MPEG-4 standard. Using this technology, 3D coordinates of one particular point of the face cannot be obtained without additional computation. For this reason, we've used the calibration system made of 2 orthogonally standing internet cameras: "Creative live Camoptia" and "Creative Live! Cam Motion" (resolution 640 x 480). These cameras capture the human face from two orthogonal views: a front and a profile. A number of feature points is already located on both (2D) views and it helps to deduce the 3D positions of the head we wish to model.

Having located the set of characteristic feature points in both views, the calculation of their 3D coordinates is carried out using perspective projection camera system, which is shown in Figure 3.

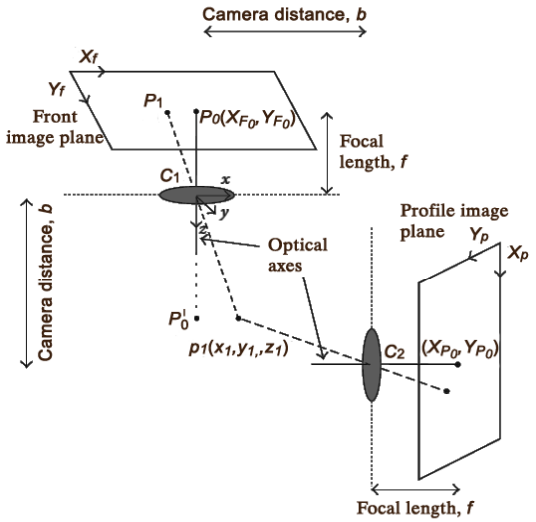


Figure 3. Image acquisition layout

The points in 3D space with coordinates (*x, y, z*) are projected on two image planes, the front and the profile, with perspective rays passing through the two corresponding projection centres *C*1 and *C*2, which lie within the physical camera and are at a distance *b* from each other [7]. *f is* the focal length. The projection coordinates for the image got from the frontal (X_{F_0}, Y_{F_0}) and profile (X_{P_0}, Y_{P_0}) cameras can be found as:

$$X_F = f\frac{-x}{z} + X_{F_0}, \quad Y_F = f\frac{-y}{z} + Y_{F_0},$$

$$X_P = f\frac{b-z}{b-x} + X_{P_0}, \; Y_P = f\frac{-y}{b-x} + Y_{P_0}. \quad (1)$$

The 3D position of the feature point (*x, y, z*) is determined by least squares method.

4.2. Speaker and text material

The drawback of used marker technique is that markers may temporally disappear, e. g. the markers on the lips may not be visible during a bilabial closure and the markers on the lower lip may be covered when the lips are protruded. To eliminate this disadvantage, the experiment was repeated five times at the same conditions (the surrounding environment was silent and well enlightened). The speaker was male and a native Lithuanian speaker. We captured only one person for our experiment, because it makes the recording process and the interpretation of the data simpler than if several speakers were to be recorded. Furthermore, such a database can be used to improve speech synthesis, if we would like to capture characteristics of one specific speaker.

The speech material consists of 9 Lithuanian words with known transcriptions earlier used by other researchers and 5 everyday sentences containing 50 words. Text was chosen to cover all Lithuanian phonemes [10].

The speaker was asked to hold the text parallel with his eyes and to read the text. For technical reasons the recordings were made in periods separated by pauses of 10 seconds. Video material was saved using MPEG-4 standard in .AVI format.

Despite the fact, that speaker tried not to move his head during experiment, head movements are inevitable during natural speech. To avoid error when modelling 3D viseme, every 3D point calculated earlier has to be transformed so that it would be displaced as close as possible to the coordinates of head at the silent stage. The relation between the initial and the transformed model is given by

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \mathbf{R}\begin{bmatrix} x_0 \\ y_0 \\ z_0 \end{bmatrix} + \mathbf{T}, \quad (2)$$

where (x_0, y_0, z_0) are the coordinates of a model node at its initial position, (*x,y,z*) are the coordinates of the transformed node, **R** is a 3x3 rotation matrix, and **T** is a 3D translation vector $\left[\mathbf{T}_x \mathbf{T}_y \mathbf{T}_z\right]^T$.

4.3. Data processing

Speech animation can be generated using the loaded sound file and the set of visemes defined in particular software. In order to get the realistic result, lip movements must be perfectly synchronized with the audio. There are 3 items which strongly influence the quality of animation. They are speech recognition engine, naturally looking 3D visemes and the correctness of phoneme to viseme mapping table.

The first one is speech recognition engine and its accurateness when we are trying to detect the time information when the new phoneme appears. In order to find the Lithuanian phoneme to viseme mapping table, the input sound file has to be extracted from video material of speech. The extracted acoustical speech was saved in .WAV output format (channel stereo, bit rate 128 kbps, sample rate 48000). Because the maximum length of sound file which software "Crazy Talk 5" is able to animate is limited to 30 seconds, we cut the speech file into segments smaller

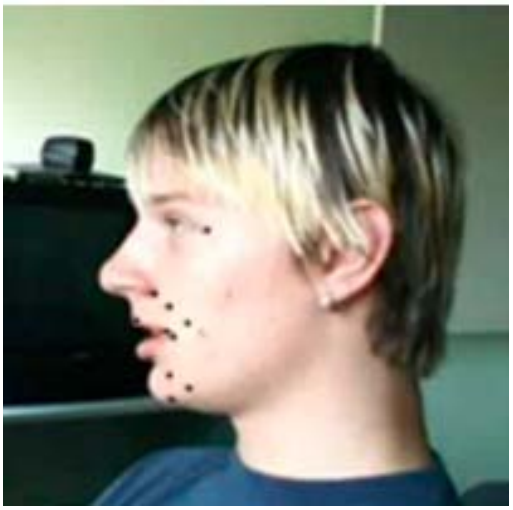
than 30 seconds. When we load the input sound into the software, phonemic alignment is done automatically. To analyze how precisely speech recognition engine (base language – English) detects phonemes positions of Lithuanian speech, we compared the automatically marked moments when the new phoneme appeared and the time information when we heard the exact phoneme.

It was detected by visual observation, that if there were no noises and the speech was expressive in the recorded file, phonemes timing of Lithuanian speech was done with the accuracy of about 90% (if the quality of sound was worse, the accuracy was about 50%). Although phonemes were detected correctly, animation was not convincible, because the animation engine used the English phoneme to viseme mapping table. We can make the conclusion that the phonemes timings are good, but the incorrect visemes are chosen when animating Lithuanian speech. The phoneme to viseme mapping table, made by visual similarity between Lithuanian and English has to be used, to change the automatically chosen viseme to the correct one.

The second important item in speech animation is naturally looking 3D visemes. The visemes integrated in the software look natural and are photorealistic, so we've used them for animation without additional edition.

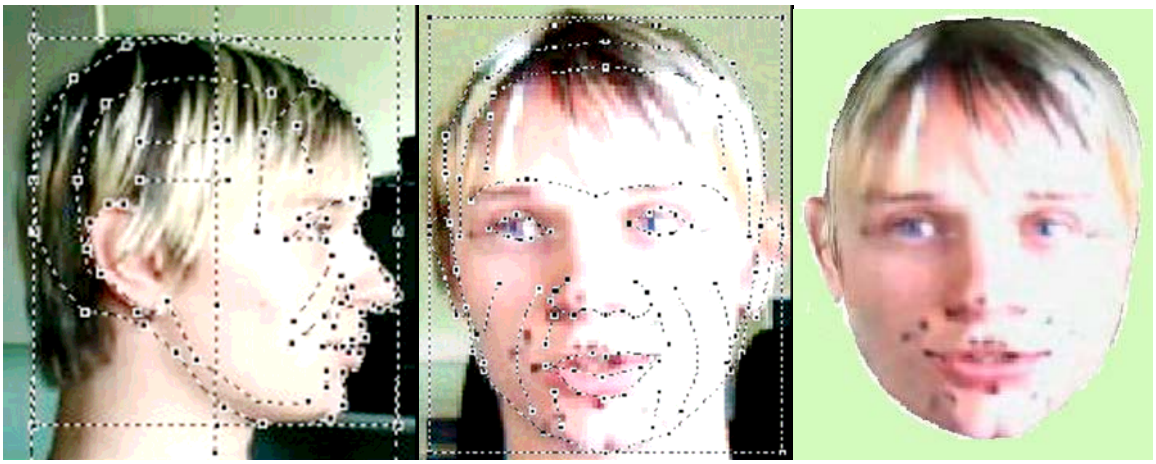
The most important thing we were concentrated on was the creation of phoneme to viseme mapping table. It was made by visual similarity between Lithuanian and English sounds. To find which English viseme matches Lithuanian sound, we had to research every detected phoneme separately. For this, 3D viseme had to be created using 2 orthogonal pictures captured at the middle of the phoneme. The sample is shown in Figure 4.

a)



b)

Figure 4. Lithuanian phoneme /ʒ/ captured: a) from the front; b) from the left of the speaker. The feature points marked according to MPEG-4 standard are seen



a) b) c)

Figure 5. Points drawn on the face fitted with the MPEG-4 feature points of a generic head: a) in the portrait picture; b) in the profile picture and c) 3D viseme of Lithuanian sound modelled using calibration of portrait and profile pictures

For the creation of Lithuanian 3D viseme we've used the modified technology of static viseme modelling [18]. First of all, using the calibration of two orthogonal pictures 3D coordinates of the feature point have to be gained. Later every principal vertex of the generic head model has to be translated to the calculated 3D position of the feature point and the 3D model of the head is generated.

To implement the technology, we've used the freeware "Faceworks" [12], which generates 3D head model using two orthogonal pictures (in profile and in portrait) of human head. This freeware was chosen due to the fact that the feature points of its generic head are arranged by MPEG-4 standard, so it was very easy to fit its feature points to MPEG-4 points drawn on the human face before capturing the head pictures. The interface and modelled 3D viseme are shown in Figure 5.

As it was mentioned earlier, accidental turns of head can't be avoided when the person talks. In order to eliminate this deviation for better visual comparison results, we've used the software function to rotate the outcome 3D head model.

Having the 3D models of 30 Lithuanian phonemes and the photorealistic English visemes, we've performed the visual comparison.

For the animation of the speech, "Crazy Talk 5" uses its own generic head model. In order to create photorealistic talking head, we had to load the portrait picture of the human face and to translate the 4 main feature points of the generic head of the "Crazy Talk 5" to the positions of lip and eye corners in the face. Later, we've had to load the explored and already created 3D Lithuanian viseme into the "Faceworks".

We visually compared every Lithuanian viseme with each of 16 English visemes (the interested English viseme appears in the software by clicking on its name in viseme table). The best match was confirmed as the relation between English and Lithuanian visemes.

Lips and jaw moving vectors vary when pronouncing different phonemes, so if more than one English viseme is similar to 3D Lithuanian viseme, speech dynamics have to be governed to get the most accurate AV mapping. For instance, lips move horizontally when we pronounce the phoneme /a/, but more vertically then the phoneme /a:/ is the object of research. Considering the vector magnitude of the movement of the lips feature point, the same English viseme can be matched as different Lithuanian viseme. In our research, speech dynamics was governed by parameter of the expressiveness and its value for every phoneme is shown in the viseme to phoneme mapping table.

The visual comparison of Lithuanian phoneme /ʒ/ is shown in Figure 6. As we see, Lithuanian 3D viseme of the phoneme /ʒ/ is very similar to English viseme "Ih" (expressiveness parameter - 100), but the more accurate mapping was defined between Lithuanian viseme /ʒ/ and English viseme "Ch_J" (expressiveness parameter – 80).

According to the results, English viseme and its matching Lithuanian phonemes written using SAMPA standard [19] were related in the Lithuanian phoneme to English viseme mapping table (Table 1). Parameter of expressiveness and Lithuanian letter representing the phoneme were recorded in the table too.

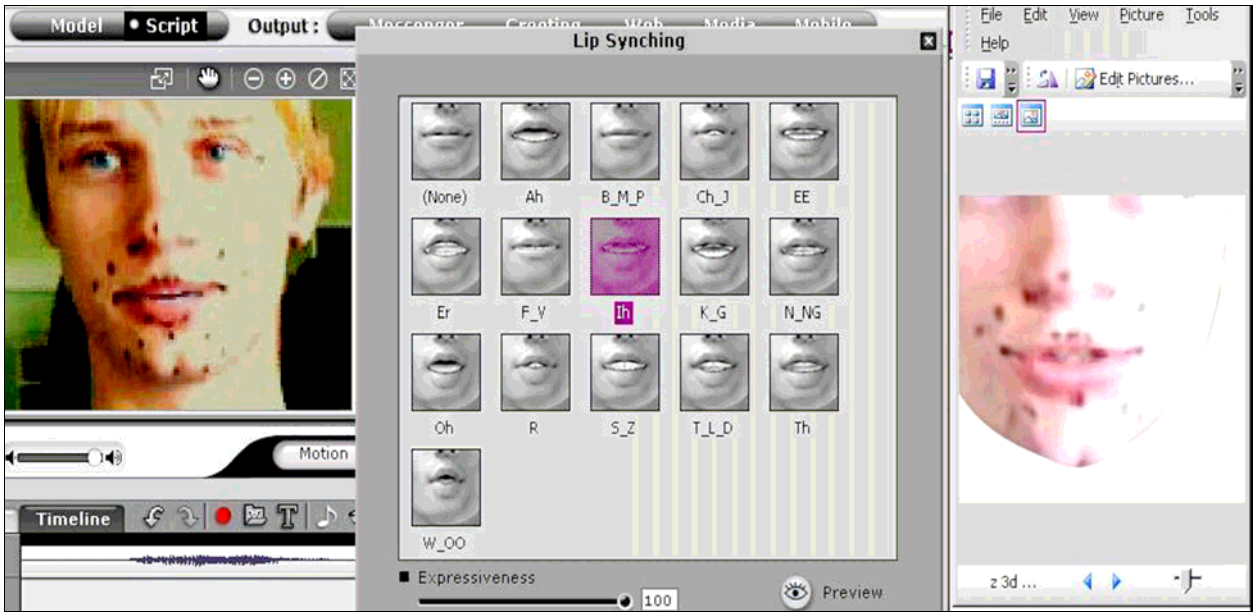


Figure 6. Visual comparison of Lithuanian 3D viseme of the phone /ʒ/ (on the right) and the English viseme /Ih/ (on the left) used to fill phoneme to viseme mapping table. Viseme /Ih/ was chosen, because it is very similar to mouth shape when pronouncing Lithuanian letter Ž

1 Table. Lithuanian phoneme to English viseme mapping table

English viseme	Lithuanian phoneme, parameter of expressiveness, corresponding Lithuanian letter
None	/silence/ 50
Ah	/a/ 80 (A), /e/ 50 (E), /e:/ 90 (Ę)
B_M_P	/b/ 80 (B) , /m/ 80 (M) , /p/ 80 (P)
Ch_J	/tʃ/ 100 (Č), /ʃ/ 100 (Š), /ʒ/ 80 (Ž)
EE	/i:/ 100 (Į, Y)
Er	/ė:/ 80 (Ė), /r/ 100 (R)
F_V	/f/ 80 (F), /v/ 80 (V)
Ih	/i/ 50 (I)
K_G	/g/ 80 (G), /k/ 80 (K)
N_NG	/n/ 100 (N)
Oh	/a:/ 80 (Ą), /o/ 60 (O)
S_Z	/ts/ 80 (C), /s/ 80 (S), /z/ 80 (Z)
T_L_D	/d/ 90 (D), /l/ 100 (L), /t/ 100 (T)
Th	/j'/ 100 (J)
W_OO	/u/ 50 (U), /u:/ 100 (Ū)

5. Conclusions

The existing speech recognition and animation engine with base language – English was adapted to animate recorded Lithuanian speech. The transmission between two languages was realized using Lithuanian phoneme to English viseme mapping table presented in this paper. The visual similarity of 30 basic Lithuanian visemes and the photorealistic English visemes was taken into consideration to classify Lithuanian phonemes into 15 English viseme classes (including silence). Lithuanian 3D visemes were modelled using the visual information of 30 basic Lithuanian phonemes and orthogonal image calibration technology.

The generated phoneme to viseme mapping table was successfully applied in the animation software „Talking Head" by manually replacing automatically wrongly identified visemes with the right ones. So the significant result of our experiment is that we don't need to build a new speech recognition and animation engine to animate Lithuanian speech. It can be easily done using English speech animation engine and presented Lithuanian viseme to English phoneme mapping table.

By the opinion of three especially not trained observers, Lithuanian speech animation looked quite realistic.

Our next step is to explore all Lithuanian phonemes including diphones and to upgrade the phoneme to viseme mapping table.

References

[1] **J. Beskow.** Rule-based visual speech synthesis. *Proceedings of the 4th European Conference on Speech Communication and Technology* (*Eurospeech'95*), *Madrid, Spain*, 1995, 299-302.

[2] **E. Bevacqua, C. Pelachaud.** Expressive audio-visual speech. *Journal of Visualization and Computer Animation*, 2004, *Vol.*15, *No.* 3-4, 297−304.

[3] **M. Brand**. Voice Puppetry. *Proceedings of the 26th annual conference on Computer graphics and interactive techniques. ACM Press/Addison-Wesley Publishing Co., New York*, 1999, 21-28.

[4] **P. Cosi, C.E. Magno, G. Perlin, C. Zmarich**. Labial coarticulation modeling for realistic facial animation. *Proceedings of 4th International Conference on Mutlimodal Interfaces, Pittsburgh, PA, USA*, 2002, 505–510.

[5] **Zh. Deng, J. Y. Noh.** Computer facial animation: A survey. *Data-Driven 3D Facial Animation. Springer Press, London*, 2007, 13-19.

[6] **T.A. Faruquie, C. Neti, N. Rajput, L.V. Subramaniam, A. Verma.** Translingual visual speech synthesis. *International Conference on Multimedia and Expo*, *New York*, 2000, *July-August, Vol.*2, 1089-1092.

[7] **N. Grammalidis, N. Sarris, F. Deligianni, M.G. Strintzis**. Three-Dimensional Facial Adaptation for MPEG-4 Talking Heads. *EURASIP Journal on Applied Signal Processing, Special Issue on Signal Processing for 3D Imaging and Virtual Reality*, 2002, *No.*10, 1005-1020.

[8] **F.J. Huang, T. Chen**. Real-time lip-synch face animation driven by human voice. *Proceedings of IEEE Multimedia Signal Processing Workshop, Los Angeles, California*, 1998, 352-357.

[9] **I. Karlsson, A. Faulkner, G. Salvi.** SYNFACE – A Talking Face Telephone. *8th European Conference on Speech Communication and Technology, Geneva, Switzerland*, 2003, 1297-1300.

[10] **P.Kasparaitis.** Lithuanian Speech Recognition Using the English Recognizer. *Informatica*, *Vol.*19, *No.*4, 2008, *December*, 505-516.

[11] **S.A. King, R.E. Parent**. Creating speech-synchronized animation. *IEEE Transactions on Visualization and Computer Graphics*, 2005, *Vol.*11, *No.*3, 341-352.

[12] **LOOXIS GmbH. Faceworx.** 3D head out of two standard 2D photos creation technology. *June* 2009, *http://www.looxis.com/en/k75.Downloads_Bits-and-Bytes-to-download.htm*.

[13] **D. W. Massaro, J. Beskow, M. M. Cohen, C. L. Fry, T. Rodriguez**. Picture my voice: audio to visual speech synthesis using artificial neural networks. *Proceedings of Auditory-Visual Speech Processing* (*AVSP* '99), *Santa Cruz*, 1999, 133-138.

[14] **A. Moura, I. Mazonaviciute, J. Nunes, J. Grigaravicius.** Human lips synchronisation in Autodesk Maya. *Systems, Signals and Image Processing* 2007 *and 6th EURASIP Conference focused on Speech and Image Processing, Multimedia Communications and Services*, 2007, 365 – 368.

[15] **T. Ohman.** An audio-visual speech database and automatic measurements of visual speech. *Quarterly Progress and Status Report, Department of Speech, Music and Hearing, Royal Institute of Technology*, 1998, *Vol*.1-2, 61-76.

[16] **J.J. O´Neill.** Contributions of the visual components of oral symbols to speech comprehension. *Journal of Speech and Hearing Disorder*, 1954, *Vol*.19, 429-439.

[17] **I.S. Pandic, R. Forchheimer.** MPEG-4 Facial Animation – The Standard, Implementation and Applications. *John Wiley & Sons Ltd*, 2002.

[**18]** **F. Pighin, J. Hecker, D. Lischinski, R. Szeliski, D. H. Salesin.** Synthesizing Realistic Facial Expressions from Photographs. *Proceedings of the* 25*th Annual Conference on Computer Graphics*, *Orlando, FL, USA*, 1998, *July* 19-24, 75-84.

[19] **A. Raskinis, G. Raskinis, A.Kazlauskiene.** SAMPA (Speech Assessment Methods Phonetic Alphabet) for Encoding Transcriptions of Lithuanian Speech Corpora. *Information Technology and Control, Kaunas*, 2003, 50-56.

[20] **Reallusion Inc. Crazytalk**. Speech animation studio. *June* 2009, *http://www.reallusion.com/crazytalk/ct_training_manual.html*.

[21] **G. Zoric.** Real-time Animation Driven by Human Voice. *Proceedings of the* 7*th International Conference on Telecommunications ConTEL*2003, *Zagreb*, 2003, 733-735.

Received January 2009.