FRAMEWORK FOR LITHUANIAN SPEECH ANIMATION

Ingrida Mazonaviciute, Romualdas Bausys

Department of Graphical Systems, Vilnius Gediminas Technical University
Sauletekio al. 11. SRKII - 608, LT-10223, Vilnius, Lithuania
phone: + (370) 52744848, fax: (+370) 52744844, email: ingrida.mazonaviciute@vgtu.lt
web: www.vgtu.lt

ABSTRACT

Speech animation is widely used in technical devices to allow the growing number of hearing impaired persons, children, middle-aged and elderly equal participation in communication. This article presents a framework, which can be used to animate Speech sound file. The architecture proposed is made up of the following components: Lithuanian speech recognition, Lithuanian 3D visemes usage according to the MPEG-4 standard and finally translingual viseme and phoneme synthesis. All the components are integrated in an English visual speech synthesizer iFACE for creation of animated Lithuanian speech samples.

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. Therefore animated characters like talking heads with lip shape mapping to specific synthesized or natural speech, are playing the considerably important role in human computer communication. Despite of the importance of synthetic speech animation in movie, advertising and computer game industries, talking heads can be widely used for interactive applications, where User Interface agents can be developed for the application in e-learning, Web navigation or as virtual secretary [1]. 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.

Talking heads are mostly driven by English phonetics (or visemes). Recently we also see talking heads driven by Finnish [2], Italian [3], Chinese Mandarin (Putonghua) and Cantonese [4]. One of the promising approaches is to integrate a talking head based on phonetics in one language, with input audio speech in another (target) language.

The goal of this paper is to propose speech animation architecture suitable for Lithuanian Speech Animation. Data flows are going to be integrated in free/open source facial animation framework compatible with MPEG-4 standard called iFACE [5]. The basic input speech language of this framework is English, so we'll suggest the architecture how it may be driven by input Lithuanian speech. For this reason both visual and acoustical aspects of Lithuanian speech will be explored.

This paper is organized as follows: the 2nd section presents face animation scripting languages and frameworks. In the Section 3 analysis of the aspects of translingual phoneme to viseme mapping is presented. Section 4 describes the architecture that we have proposed for development of the framework for Lithuanian speech animation.

2. SCRIPTING LANGUAGES FOR FACIAL ANIMATION

For the generation of naturally speaking talking head, positions of the mouth and tongue must be related to characteristics of the speech signal. Many recent researches [6,7,8,9] on facial animation are focused on a facial animation platform development. Face parameterization and scripting languages for assisted and automatic animation generation are also widely discussed. One of the most popular standards for parameters set is MPEG-4 Facial Animation (MPEG-4 FA) [10], which includes Face Definition Parameters (FDPs) and Face Animation Parameters (FAPs). FDPs define a face by giving measures for its major parts and features such as eyes, lips and their related distances. On the other hand, FAPs encode the movements of these facial features. Together they allow for a receiver system to create a face (using any graphical system) and animate that face by a set of low level commands in terms of FAPs. However, although MPEG-4 FDPs and FAPs are powerful means in facial animation, they do not provide an animation language but only a set of low-level parameters. The content providers and animation engines still needed higher levels of abstraction on the top of MPEG-4 parameters to provide group actions, timing control, event handling and similar functionality usually provided by a high-level language [5].

The absence of a dedicated language for a facial animation has been evident especially within the XMT framework. Recent advances in developing and using Embodied Conversational Agents, especially their web-based applications, and growing acceptance of XML as a data representation language, have drawn attention to mark-up languages for virtual characters. The basic idea was to define specific XML tags related to agents' actions such as speech production, facial and body animation, emotional representation, dialogue management etc.

 ISSN 2076-1465

	SMIL	VRML	FAML	MPML	BEAT	FML
Face-specific Parts	No	No	Yes	Yes	Yes	Yes
MPEG-4 Compatible	No	No	No	No	No	Yes
Timing Control	Yes	Partial	Yes	Yes	Yes	Yes
Decision-making	Yes	Partial	No	Yes	Partial	Yes
XML-based	Yes	Yes	Yes	Yes	Yes	Yes
High-level Face Components	No	No	Partial	Partial	Partial	Partial
Behavioural Modelling	No	No	No	No	Yes	Partial

Table1 – Content description methods and facial animation features.

Table 1 summarizes major languages that may be used for facial animation and their supported features. The need for a unifying language specifically designed for facial animation that works as an abstraction layer on top of MPEG-4 parameters was the main motivation in designing Face Modelling Language (FML) which allows re-use of existing XML tools and products [11].

Face parameterization and scripting languages are used to develop facial animation frameworks, which aim to provide a desirable platform for facial animation research or the development of talking heads applications [12].

Wang [6] described a methodology for the construction of an expressive facial animation system with lip synchronization. He used affordable off-the-shelf components which are provided by the FaceGen Modeller software for face key meshes generation and the Microsoft Speech SDK as the speech API.

Cosi [7] proposed a facial animation toolkit implemented in MATLAB created mainly to speed up the procedure for building the LUCIA talking head through motion capture techniques, translated to MPEG-4 parameters.

Balci et al. [8] designed Xface, a set of open source tools for creation of talking heads using MPEG-4 and keyframe based animation. Xface uses the SMIL-Agent scripting language for its keyframe-based animation module. A set of key-meshes with the different facial expressions and visemes must be defined for each talking head model.

DiPaola and Arya [9] proposed a facial animation framework compatible with the MPEG-4 standard called iFACE. iFACE allows interactive non-verbal scenarios through FML scripting language [5]. FML allows both parallel and sequential description of face actions that include talking, expressions, head movements and low-level MPEG-4 parameters.

Our approach also uses iFACE as the facial animation engine and proposes integration architecture for different speech animation components.

3. TRANSLINGUAL PHONEME TO VISEME MAPPING

The speech recognition and animation engine is a critical part of any speech animation system. 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 we assume that Lithuanian is the novel language and English is the base language [13].

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. Lithuanian and English phonemes are related according to the table proposed by Kasparaitis [14].

Many acoustic sounds of separate languages are visually similar and accordingly different phonemes can be classified using the same viseme. Since the system iFACE has to work for novel Lithuanian 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. Lithuanian phoneme to English viseme mapping table [13] are going to be employed for translingual phoneme to viseme mapping. Since the table describes 30 of 58 Lithuanian phonemes and some visemes differ from 15 static visemes defined in MPEG-4 standard, expansion of this table should be performed. First of all, we need to append the Lithuanian phoneme to English viseme mapping table with the column defining the numbers of the English visemes in MPEG-4 standard (the fragment of this is shown in Table 2). It will simplify Lithuanian phonemes translingual definition in FML.

Viseme number (by MPEG-4)	English viseme	Lithuanian phoneme, parameter of expressiveness, corresponding Lithuanian letter
0	None	/silence/ 50
1	B M P	/b/ 80 (B) , /m/ 80 (M) , /p/ 80 (P)
2	F V	/f/ 80 (F), /v/ 80 (V)
3	Th	/j'/ 100 (J)
10	Oh	/a:/ 80 (Ą), /o/ 60 (O)
11	Ah	/a/ 80 (A), /e/ 50 (E), /e:/ 90 (Ę)

Table 2 – Fragment of Lithuanian phoneme to English viseme mapping table augmented by viseme number in MPEG-4 standard.

However, some Lithuanian sounds (like Į, Y, Č, Ė etc.) don't have a corresponding viseme defined in MPEG-4 standard at all. Additional research must be performed to group all the Lithuanian phonemes with respect to their visual similarity and to create Lithuanian phoneme to Lithuanian viseme mapping table. Nevertheless, the generation of Lithuanian phoneme to Lithuanian viseme mapping table is not the goal of this paper, it's the area of future researches, so we don't propose the generation details here.

4. FRAMEWORK FOR LITHUANIAN SPEECH ANIMATION

Talking heads can be driven by input text or input speech. While text-driven talking heads employ both synthesized voices and head models, constituting text-to-audiovisual speech; speech-driven talking heads involve synthesizing visual speech information from genuine speech that makes animation more realistic, but more complicated to produce, Speech recognition is still the one of the most challenging areas for researches.

For the generation of lip animation, we use iFACE as the engine for implementing the face object within multimedia systems. Naturally, iFACE uses its own phoneme speech alignment tool, which comes with HTK 2.0 for phoneme recognition and alignment. Speech stream is decoded into phoneme sequence with duration information. The integrated English speech recognition engine doesn't produce satisfactory results for the Lithuanian speech, when we're trying to get syllable transcription and the timing information. Thus, we propose to utilize Lithuanian speech recognition engine which was developed by Lithuanian speech recognition researchers team consisted of Lipeika and etc. [15].

The overall architecture of our framework designed to animate Lithuanian speech is presented in Figure 2. The data flow is organized as follows:

1. Firstly, phonetic transcription and the timeline of the Lithuanian phonemes (Figure 1) are constructed by Lithuanian speech recognition engine. Lithuanian speech sound file (.wav) is the input for the engine.
2. The complete timeline of visemes is generated by translingual phoneme to viseme mapping module. Lithuanian language alignments (Figure 1), Lithuanian phoneme to English viseme mapping table (Table 2) and the additional Lithuanian phoneme to Lithuanian viseme mapping table are merged to perform translingual mapping, which connects the phonemes to their visual representation (visemes).
3. The linear interpolation is applied for speech synchronization between FAPs of two adjacent visemes in the timeline. English and additional Lithuanian visemes are defined by FAP parameters and stored in the separate file. Speech phonetic transcription (Figure 1) is described by FML scripting language specified to describe facial actions. Finally, Lithuanian speech audio file (.wav), 3D geometry head file (.msh) (editable in geometry and texture) and the animation script (.fml) are compound into iFACE to get video of Lithuanian talking head synchronized with speech.

```
File type = "ooTextFile"
Object class = "TextGrid"
xmin = 0
xmax = 4.2860090702947851
tiers? <exists>
size = 3
item []:
    item [1]:
        class = "IntervalTier"
        name = "qarsai"
        xmin = 0
        xmax = 4.2860090702947851
        intervals: size = 25
        intervals [1]:
            xmin = 0
            xmax = 0.20000000000000001
            text = "..."
        intervals [2]:
            xmin = 0.20000000000000001
            xmax = 0.3387614689065202
            text = """a"
        intervals [3]:
            xmin = 0.3387614689065202
            xmax = 0.46149769109181304
            text = "k"
        intervals [4]:
            xmin = 0.46149769109181304
            xmax = 0.62609926758129486
            text = "t'"
        intervals [5]:
            xmin = 0.62609926758129486
            xmax = 0.81475200567350736
            text = "i"
```

Figure 1 – Phonetic transcription and the timeline of Lithuanian word "akti"".

4.1 iFACE as the platform for speech animation

iFACE (Interactive Face Animation – Comprehensive Environment) facial animation engine is the base element for the proposed Lithuanian speech animation framework. iFACE uses Microsoft Direct3D, DirectSound and .NET frameworks to allow interfacing through web services and other distributed components. Its Stream Layer components are built on the basis of DirectShow technology in order to use the built-in streaming functionality.

iFACE was chosen for the following reasons:

1. Hierarchical 3D head model for controlling facial actions from vertex to feature-group levels are incorporated in it. The adjustment can be simply done by editing vertices and attaching new textures to it. Also, iFACE supplies the possibility to import 3D head mesh through .msh file, so it is easy to integrate a new head model.
2. The synchronization of speech acoustic and visual output can be easily performed by iFACE. Audio kernel processes an input audio data and at the same moment the timeline of events and actions described in FML file goes through the video kernel in order to create video frames corresponding to desired facial actions. Possibility to model behavioural logic, when actions of an agent (similar to people) is based on stimulus-response model, are also very important for realistic speech animation.
3. FML is the scripting language of iFACE system. FML characteristics like MPEG-4 compatibility, timing control, decision-making, possibility to characterize distinctive visemes by MPEG-4 FAP parameters (Table 1) allow us to create interactive animation scenarios.

These three features made iFACE system as the attractive basis for Lithuanian speech animation.

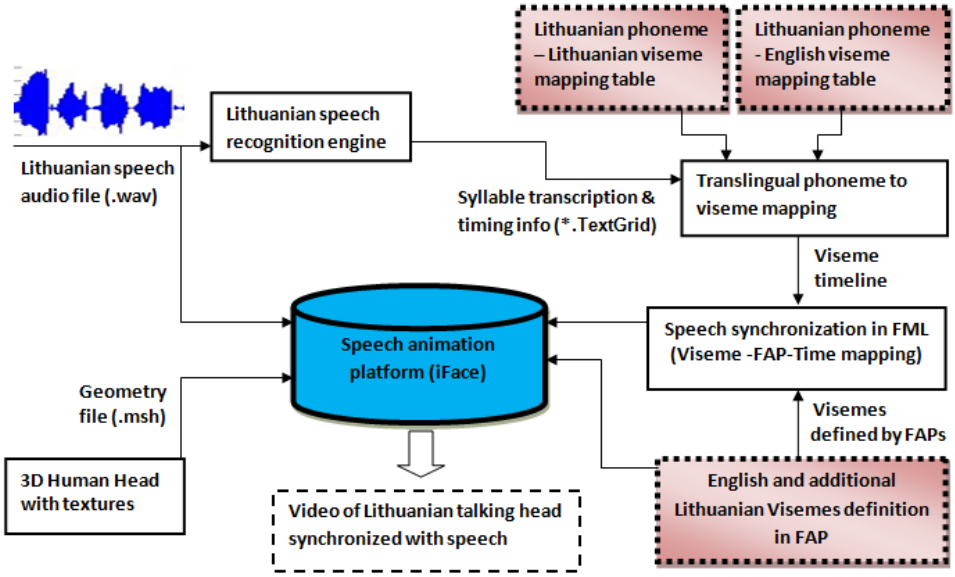

Figure 2 – Architecture of Lithuanian speech animation framework.

4.2 FML as the language for speech synchronization compatible with MPEG-4

iFACE facial animation system uses FML as the content description mechanism. FML - unifying language specifically designed for facial animation that works as an abstraction layer on top of MPEG- 4 parameters. FML compatibility with MPEG-4 (MPEG-4 FAPs are supported explicitly and FDPs implicitly by general purpose model definition mechanisms) together with XML and related web technologies guarantee that our animation scripts later could be simply used in speech animation web applications. Independence of the type of head model, timeline definition of the relation between facial actions and external events together with hierarchical representation of face animation mean that in one FML script we can define frames, simple moves, but also meaningful actions and even stories. It will be beneficial, when we'll animate long Lithuanian speeches. Simple linear interpolation between neighbouring visemes is described by FML language also.

Timing information which is obtained by Lithuanian speech recognition engine (Figure 1) goes through translingual adaptation. As we see in the animation script sample (Figure 3), visemes are defined by FAP parameters. There are 68 face animation parameters which define the deformation of character's head in MPEG-4 standard. The first two suit a framework with the high level parameters, representing visemes and the six basic emotions [12]. The next ones deal with specific regions on the face, as left eyebrow, right corner lip, tongue tip, etc.

So the code line: "<param type="FAP" name="1-1-5" value="80" begin="20" end="27" />" means, that viseme number 5 (by MPEG-4 standard) will be shown in the period between 20ms and 27ms. Additionally, expressions like sadness, joy etc. and the head movements can be described in FML script.

Having the phoneme-viseme-FAP mapping tables, FML script is processed by FML interpreter in iFACE system. The result it animated talking head synchronized with Lithuanian sound file.

```
<?xml version="1.0"?>
<fml>
  <model>
  </model>
  <story>
    <act>
        <seq event_value="2">
          <hdmv type="yaw" value="15" begin="0" end="50" />
          <param type="FAP" name="1-1-0" value="50" be-
gin="0" end="12" />
          <param type="FAP" name="1-1-15" value="80" be-
gin="12" end="20" />
          <param type="FAP" name="1-1-5" value="80" be-
gin="20" end="27" />
          <param type="FAP" name="1-1-4" value="100" be-
gin="27" end="38" />
          <param type="FAP" name="1-1-12" value="50" be-
gin="38" end="49" />
        </seq>
    </act>
  </story>
</fml>
```

Figure 3 – FML script for animation of word "akti" (translation – "to become blind").

5. SUBJECTIVE EVALUATION

Source material consisted of 3 different length Wave Sound (.wav) files. To follow the visibility of the phoneme "a" in a word, separate words were grouped into 3 categories and recorded like separate sound files. Speech records were processed through the proposed Lithuanian speech animation framework in order to generate speech video files (.avi). To evaluate the animation quality, they were presented to 15 students, who judged knowing the recorded text. Speech and

video synchronization quality was estimated considering if the visual speech looks natural, understandable and believable within a MOS scale (5 = excellent, 4 = good, 3 = fair, 2 = poor, 1 = bad).

Word categories were organized as follows:

CASE **A**. Entirely English phonemes, short /a/ *(""""akt'i", "k""ap's'i", "kas""a", "r""azdavo:"*). This group of recorded words gave the worst result for speech animation (MOS – 1,5). The observers mentioned that short /a/ in a short word is hardly seen and they roughly undestanded the visual speech.

CASE **B**. English phonemes and additional Lithuanian phonemes, long words, long /a:/ corresponding Lithuanian letter - "ą" *(""""a:Zuolas", "prak""a:sto:"),*

CASE **C**. Entirely English phonemes, long English /a:/ *("^a:s'ilas", "k^a:pas", "b^a:das").*

Cases **B** and **C** showed better results (case B – MOS 3, case C – MOS 3). Cases B and C were better understandable because the visual shape for long /a:/ is much more expressive than the one for the short /a/. Observers mentioned that video of Case **B** was more recognizable when in CASE C because words were longer (3 sylablles).

On the basis of numerical experiments it is possible to conclude, that speech animation would be better understandable if analyzed words would be included in a sentence. Also, that emotions would improve the perception of the animated speech.

6. CONCLUSIONS

Speech animation architecture suitable for Lithuanian Speech Animation was proposed in this paper. Facial animation framework compatible with the MPEG-4 standard called iFACE was chosen to integrate different speech animation components. The basic input speech language of the framework is English. Translingual phoneme to viseme mapping technology was applied to force iFACE to animate recorded Lithuanian speech. Linear interpolation between neighbouring visemes was performed to define visual speech coarticulation by Face Modelling Language. Lithuanian speech recognition engine, translingual phoneme to viseme mapping tables, 3D hierarchical head model, 3D Lithuanian visemes defined by FAP parameter and FML script were merged to get video of Lithuanian talking head synchronized with speech. Numerical experiments performed to valuate the visual comprehension of speech were performed. Although additional speech synchronization and coarticulation rules must be integrated and described by FML language, proposed architecture is presentable for Lithuanian speech animation.

REFERENCES

[1] A. Moura, I. Mazonaviciute, J. Nunes and 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, pp. 365 – 368.

[2] J. L. Olives, M. Sams, J. Kulju, O. Seppaia, M. Karjalainen, T. Altosaar, S. Lemmetty, K. Toyra and M. Vainio, "Towards a High Quality Finnish Talking Head", *IEEE 3rd Workshop on Multimedia Signal Processing*. 1999, pp. 433–437.

[3] C. E. Pelachaud, Z. Magno-Caldognetto and P. Cosi, "Modelling an Italian Talking Head". *In Proc. Audio-Visual Speech Processing*. 2001, pp. 72–77.

[4] J. Q. Wang, K. H. Wong, P. A. Heng, H. Meng and T. T. Wong, "A Real-Time Cantonese Text-To-Audiovisual Speech Synthesizer", in *Proc. ICASSP 2004*. 2004, pp. 653–656.

[5] A. Arya and S. DiPaola, "Face Modeling and Animation Language for MPEG-4 XMT Framework", *IEEE Transactions on Multimedia*, vol. 9, No. 6, pp. 1137-1146. 2007.

[6] A.Wang, M. Emmi and P. Faloutsos, "Assembling an expressive facial animation system", in *Sandbox '07: Proceedings of the 2007 ACM SIGGRAPH symposium on Video games*, New York, USA. 2007, pp. 21–26.

[7] P. Cosi, C. Drioli, F. Tesser and G.Tisato, "Interface toolkit: a new tool for building IVAs", in *Intelligent Virtual Agents Conference (IVA'05)*, Greece. 2005, pp. 75–87.

[8] K. Balci, E. Not, M. Zancanaro and F. Pianesi, "Xface open source project and smil-agent scripting language for creating and animating embodied conversational agents", in *MULTIMEDIA'07: Proceedings of the 15th international conference on Multimedia*, New York, USA. 2007, pp. 1013–1016.

[9] S. Dipaola and A. Arya, "A framework for socially communicative faces for game and interactive learning applications", in *Future Play '07: Proceedings of the 2007 conference on Future Play*, New York, USA. 2007, pp. 129–136.

[10] I. S. Pandzic and R. Forchheimer, *MPEG-4 Facial Animation: The Standard, Implementation and Applications*. England: John Wiley & Sons, 2002.

[11] A. Arya and S. DiPaola, "Face As A Multimedia Object," *in Proc. 5th International Workshop on Image Analysis for Multimedia Interactive Services*, Lisbon, Portugal, April 21-23. 2004, pp. 21 – 23.

[12] R. Queiroz, M. Cohen and S. R. Musse, "A facial animation interactive framework with facial expressions, lip synchronization and eye behavior". In *Proceedings of SBGames'08: Computing Track*, Belo Horizonte, Brazil, November. 2008, pp. 151–158.

[13] I.Mazonaviciute and R.Bausys, "English Talking Head Adaptation for Lithuanian Speech Animation" *Information Technology And Control*, vol. 38, No. 3, pp. 217 - 224. 2009.

[14] P. Kasparaitis, "Lithuanian Speech Recognition Using the English Recognizer", *Informatica*, vol.19, No.4, pp. 505-516, Dec. 2008.

[15] A. Lipeika, J. Lipeikiene and L. Telksnys, "Development of isolated word speech recognition system", *Informatica*, vol. 13, No. 1, pp. 37-46. 2002.