ELECTRONICS AND ELECTRICAL ENGINEERING
ISSN 1392 – 1215 2011. No. 5(111)
ELEKTRONIKA IR ELEKTROTECHNIKA

SYSTEM ENGINEERING, COMPUTER TECHNOLOGY
T 120
SISTEMŲ INŽINERIJA, KOMPIUTERINĖS TECHNOLOGIJOS

Translingual Visemes Mapping for Lithuanian Speech Animation

I. Mazonaviciute, R. Bausys
Department of Graphical Systems, Vilnius Gediminas Technical University,
Saulėtekio av. 11, 2608 kab., 10223 Vilnius, Lithuania, phone: +370 527 448 48, e-mail: ingrida.mazonaviciute@vgtu.lt

Introduction

Speech perception refers to the processes by which humans are able to interpret and understand the sounds used in language. Watching at lips and tongue movements of the speaker significantly improves the understanding of acoustic signal [l]. Computers generated 3D human head models, specified to animate synthesized or natural speech, are called Talking heads. They are playing considerably important role in human-computer communication and have caused significant scientific, technological and artistic interests in computer facial animation. Talking heads can be employed for e-consulting services: virtual secretary, WEB navigator or virtual agent who is responsible for information conveying to user in a Smart Ecological and Social Apartments (SESA) [2]. Also they are widely used in e-learning technologies as animated 3D models for the correct sound pronunciation presentation [3] or applied in movie, advertising and computer game industries. Most of the existing models are dedicated to animate English language, meanwhile there are proposed Talking heads for other languages [4, 5].

Talking head is driven by speech phonetics and their visual representation – visemes. Some phonemes (vowels, consonants, diphones) constantly repeat in different languages or dialects (e.g. consonants /p/, /t/, /k/, /m/, /n/). Therefore, one of the promising approaches of effective translingual speech animation is to integrate a talking head based on phonetics in one language, with input audio speech in desired language [6], since the invention of new speech animation engine is time consuming task and also requires considerable financial support and specific knowledge.

The purpose of the present research is to enlighten new aspects of translingual visemes mapping and include them into the speech animation architecture suitable for Lithuanian Speech Animation. Free/open source facial animation framework compatible with MPEG-4 standard is called iFACE [7] will be used to integrate data flows. It is basically driven by English phonetics, so we propose the architecture how it can be driven by input Lithuanian speech.

Framework for Lithuanian speech animation

Talking heads can be driven by input text or input speech. Text-driven Talking heads employ synthesized voices and synthesized head models to represent text-to-audiovisual speech, while speech-driven models utilize acoustics (and phonetic alignment) of natural human speech. Speech-driven Talking head makes language animation more realistic, but more complicated to produce, since speech recognition is still the one of the most challenging areas for researches.

We've chosen iFACE (Interactive Face Animation – Comprehensive Environment) as the facial animation engine for implementing face object within multimedia systems. Its phoneme speech alignment tool comes with HTK 2.0 for phoneme recognition and alignment. Naturalness of animation in speech-driven talking head strongly depends on language phonetics. iFACE was originally created to animate English language, that's why it doesn't produce satisfactory results when we're trying to get automatically generated syllable transcription and timing information of the recorded Lithuanian speech. Transcriptions of Lithuanian words so far is the only solution to employ popular foreign voice servers applications for Lithuanian language [8]. Before the investigation of transcriptions the suitable recognition engine for Lithuanian speech recognition has to be found. Lithuanian speech recognition engine, which was developed by researchers' team, consisted of Lipeika and etc. [9] was utilized to get phonemes arrangement in the timeline, although there were other possibilities, e.g. [10].

The overall architecture of our proposed framework designed to animate Lithuanian language is presented in Fig.1. The data flow is organized as follows [11]:

1. Firstly, phonetic transcription and the timeline of the Lithuanian phonemes 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 after the alignment of Lithuanian phonemes and their timing information goes through translingual phoneme to viseme mapping module, which is later proposed in this paper.

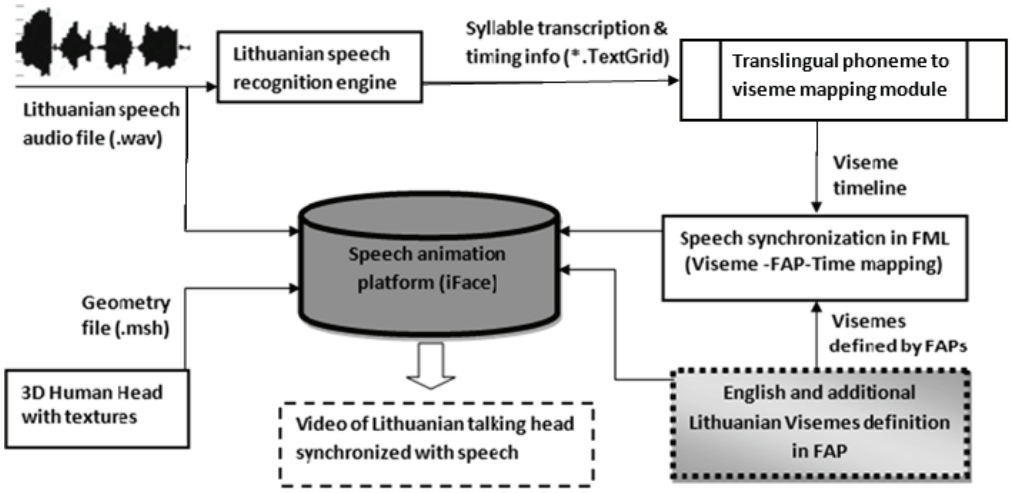

Fig. 1. Architecture of Lithuanian speech animation framework

3. Speech synchronization is described by FML scripting language, which was specified to describe facial actions [12]. Phonemes and their corresponding visemes are aligned in the timeline by FML script.

4. Finally, Lithuanian speech audio file (.wav), 3D geometry head file (.msh) (editable in geometry and texture), English and additional Lithuanian visemes defined in FAP together with the animation script (fml) are compound to get video of Lithuanian speech.

iFACE as the platform for speech visualization

iFACE 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. Three main features of iFACE made the system as the attractive basis for Lithuanian speech animation:

1. Hierarchical 3D head model for controlling facial actions. Adjustment of the model can be simply done by editing vertices and applying new textures. Also, it is easy to integrate a new head model through .msh file.

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 goes through the video kernel in order to create video frames corresponding to desired facial actions. Possibility to model behavioral logic, when actions of an agent (similar to people) are based on stimulus-response model, are also very important for realistic speech animation.

3. FML (Facial modeling language) is the scripting language of iFACE. It's compatibility with MPEG-4 together with XML and related web technologies guarantee that animation scripts 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 facial animation mean that in one FML script we can define frames, simple moves, meaningful actions and even stories. Animation script sample of Lithuanian word "akti" (translation – "to become blind"):

```
<?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" begin="0" end="12" />
          <param type="FAP" name="1-1-15" value="80" begin="12" end="20" />
          <param type="FAP" name="1-1-5" value="80" begin="20" end="27" />
          <param type="FAP" name="1-1-4" value="100" begin="27" end="38" />
          <param type="FAP" name="1-1-12" value="50" begin="38" end="49" />
        </seq>
      </act>
    </story>
</fml>
```

The code line: "<param type="FAP" name="1-1-5" value="80" begin="20" end="27" />" means, that viseme number 5 will be shown in the period between 20ms and 27ms. Phoneme duration times were obtained by Lithuanian speech recognition engine. Visemes are defined by MPEG-4 FAP parameters. Simple linear interpolation between Facial Animation Points (FAPs) of two adjacent visemes was applied to describe speech coarticulation. Expressions like sadness, joy etc. and head movements can be included in the script additionally.

Lithuanian phonemes to translingual visemes mapping

Many acoustic sounds of separate languages are visually similar and accordingly different phonemes can be classified using the same viseme. A viseme is a representational unit used to classify sounds in the visual domain. Proper visemes definition and their correct alignment are the crucial points for believable language animation, while speech recognition and animation engine is the most important part of any speech animation system. As mentioned earlier, possibility to reuse already existing speech animation engine of the base language in order to animate the novel language is very important for the

creation of speech animation applications. Base language is the language used in training the speech recognition system and the novel one is the language in which the video has to be synthesized. In this paper we use English as the base language and Lithuanian as the novel one.

According to Lithuanian grammar rules standard Lithuanian alphabet consists of 32 characters and 58 Lithuanian phonemes. In the meantime, English phoneme set is smaller and consists of 48 phonemes (the count varies). Since iFACE employs English phonetics alignment generator with viseme set and it is applied to animate Lithuanian language, direct acoustic to visual linkage (mapping) fails to produce convincing speech visualization. It was identified by testing of the framework. Thus, we must define and employ translingual phoneme to viseme mapping technique to animate Lithuanian speech. Construction of our proposed translingual phoneme to viseme mapping module is presented in Figure 2.

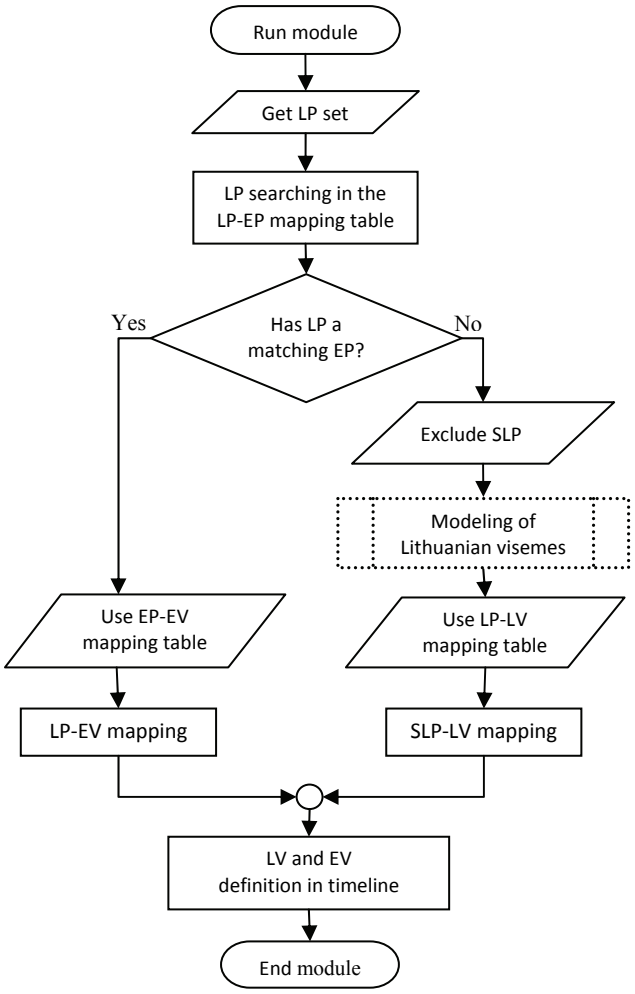

Fig. 2. Construction of translingual phoneme to viseme mapping module, when English is the base language

It is explained as follows:

1. Lithuanian Phonemes (*LP*) can be divided into 2 groups: those who have the analogous English Phoneme (*EP*) and those who don't. Lithuanian and English phonemes are related according to the *LP–EP* mapping table proposed by Kasparaitis [13]. If *LP* has a matching *EP*, we apply one of the existing English Phonemes to English Visemes (*EV*) mapping tables to find the most suitable *EV* for the analyzed Lithuanian phoneme. There are *EV*s defined in iFACE, so we propose to reuse them. Example of the *LP-EV* mapping is shown in Table 1 [13].

Table 1. Fragment of Lithuanian phoneme to English viseme mapping

Viseme No. (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	/i'/ 100 (J)
10	Oh	/a:/ 80 (Ą), /o/ 60 (O)
11	Ah	/a/ 80 (A), /e/ 50 (E), /e:/ 90 (Ę)

2. Some Lithuanian phonemes can't be associated with phonemes of the base language (e.g. Į, Y, Č, Ė). We define them as specific Lithuanian phonemes (*SLP*). In order to visualize *SLP*, we must define new visemes. If 2 or more *SLP* are visually ambiguous, they are defined as one Lithuanian viseme (*LV*). Every new *LV* has to be redefined or sculpted individually. Face expressions are defined by FAP points of MPEG-4 standard. Having *LV*s, they are linked with corresponding *LP*s in Lithuanian phonemes–Lithuanian visemes mapping table, which is used for *SLP-LV* mapping. If visemes of the novel language was already defined, we can skip the modeling step. Creation of Lithuanian visemes is described in [14].

3. Both *LP-EV* and *SLP-LV* mappings are used to set the order of visemes in the timeline, got by Lithuanian speech recognition engine.

Conclusions

There are 4 items that strongly influence the quality of speech animation: acuraccy of speech recognition engine, naturally looking 3D visemes, the correctness of phoneme to viseme mapping and speech synchronisation algorithm. Middle two of them takes advantage of the proposed translingual phoneme to viseme mapping module, which is is suitable to animate foreign languages:

1. Application of proposed module significantly reduces the amount of new visemes necessary to be defined and sculpted to animate novel language. Examination of the relation between LP and EP [13] showed that only 17% of 35 Lithuanian phonemes must be analysed for visual representation. Residual Lithuanian phonemes can be represented by English visemes .

2. The module offers possibility to reuse visemes of existing speech animation engine, although small modifications can be required: visual representations of SLP can be appended as separate units or assigned to the visemes already defined in chosen animation engine.

Translingual phoneme to viseme mapping module is part of proposed architecture of speech animation framework suitable for Lithuanian Speech Animation. Additional speech synchronization and coarticulation rules are necessary for the accuracy of animation and will be integrated later.

References

1. **O´Neill J.J.,** Contributions of the visual components of oral symbols to speech comprehension // Journal of Speech and Hearing Disorder,1954. – No. 19. – P. 429–439.
2. **Bielskis A. A., Andziulis A., Ramašauskas O., Guseinovienė E., Dzemydienė D., Gricius G.,** Multi-Agent Based E-Social Care Support System for Inhabitancies of a Smart Eco-Social Apartment // Electronics and Electrical Engineering. – Kaunas: Technologija, 2011. – No. 1(107). – P. 11–14.
3. **Pentiuc S. G., Schipor O. A., Danubianu M., Schipor M. D., Tobolcea I.,** Speech Therapy Programs for a Computer Aided Therapy System // Electronics and Electrical Engineering. – Kaunas: Technologija, 2010. – No. 7(103). – P. 87–90.
4. **Olives J. L., Sams M., Kulju J., Seppaia O., Karjalainen M., Altosaar T., Lemmetty S., Toyra K., Vainio M.** Towards a High Quality Finnish Talking Head // IEEE 3rd Workshop on Multimedia Signal Processing, 1999. – P.433–437.
5. **Pelachaud C. E., Magno–Caldognetto Z., Cosi P.** Modelling an Italian Talking Head // Proceedings of the Audio–Visual Speech Processing, 2001. – P.72–77.
6. **Wang J. Q., Wong K. H., Heng P. A., Meng H., Wong T. T.** A Real–Time Cantonese Text–To–Audiovisual Speech Synthesizer // Proceedings of ICASSP'2004, 2004. – P. 653–656.
7. **Faruquie T. A., Neti C., Rajput N., Subramaniam L. V., Verma A.** Translingual visual speech synthesis // International Conference on Multimedia and Expo. – New York, 2000. – Vol. 2. – P.1089–1092.
8. **DiPaola S., Arya A.** A framework for socially communicative faces for game and interactive learning applications //: Proceedings of the 2007 conference on Future Play '07. – New York, 2007. – P. 129–136.
9. **Rudžionis A., Maskeliūnas R., Ratkevičius K., Rudžionis V.** Investigation of voice servers application for Lithuanian language // Electronics and Electrical Engineering. – Kaunas: Technologija, 2007. – No. 6(78). – P.43–46.
10. **Laurinčiukaitė S., Lipeika A.** Syllable–phoneme based continuous speech recognition // Electronics and Electrical Engineering. – Kaunas: Technologija, 2006. – No. 6(70) – P. 91–94.
11. **Mazonaviciute I., Bausys R.** Framework for Lithuanian speech animation // Proceedings of 18th European Signal Processing Conference (EUSIPCO–2010). – Aalborg, Denmark, 2010. – P. 781–785.
12. **DiPaola S., Arya A.** Face Modeling and Animation Language for MPEG–4 XMT Framework // IEEE Transactions on Multimedia, 2007. – Vol. 9. – No. 6. – P. 1137–1146.
13. **Kasparaitis P.** Lithuanian Speech Recognition Using the English Recognizer // Informatica. – 2008. – Vol.19, No.4. – P.505– 516.
14. **Mažonavičiūtė I., Baušys R.** English talking head adaptation for Lithuanian speech animation // Information technology and control. – Kaunas University of Technology, 2009. – Vol. 38. – No. 3. – P. 217–224.

Received 2010 11 30

I. Mazonaviciute, R. Bausys. Translingual Visemes Mapping for Lithuanian Speech Animation // Electronics and Electrical Engineering. – Kaunas: Technologija, 2011. – No. 5(111). – P. 95–98.

Methodology of Lithuanian phonemes visualization using visemes set of the base language appended by new visemes defined to animate specific Lithuanian phonemes is proposed. Translingual visemes mapping for Lithuanian speech animation is applied. Phoneme to viseme mapping module is one of the most important parts of the framework for Lithuanian speech animation. Facial animation engine compatible with MPEG-4 standard is used to integrate data flows. The base language of the exploited engine is English and the proposed architecture explains how it can be driven by Lithuanian phonetics to get samples of animated Lithuanian speech. Translingual phoneme to viseme mapping module offered in this paper is suitable to animate foreign languages. Ill. 2, bibl. 14, tabl. 1 (in English; abstracts in English and Lithuanian).

I. Mažonavičiūtė, R. Baušys. Skirtingų kalbų vizemų atitaikymas lietuvių kalbai animuoti // Elektronika ir elektrotechnika. – Kaunas: Technologija, 2011. – Nr. 5(111). – P. 95–98.

Straipsnyje pristatoma lietuvių kalbos fonemų vizualizavimo metodika, kuriai naudojama bazinės kalbos vizemų aibė, papildyta specifinėms lietuvių kalbos fonemoms animuoti skirtomis vizemomis. Lietuvių kalbai animuoti naudojamas skirtingų kalbų vizemų atitaikymas. Fonemų ir vizemų atitaikymo modulis yra viena iš svarbiausių Lietuvių kalbai animuoti skirtos architektūros dalių. Duomenų srautams integruoti naudojamas veido animavimo „variklis", suderintas su MPEG-4 standartu. Jo bazinė kalba yra anglų, taigi pristatomas animavimo karkasas paaiškina, kaip „variklį" pritaikyti lietuvių kalbai animuoti. Straipsnyje pasiūlytas skirtingų kalbų fonemų ir vizemų atitaikymo modulis tinka ir kitoms pasaulio kalboms animuoti. Il. 2, bibl. 14, lent. 1 (anglų kalba; santraukos anglų ir lietuvių k.).