VILNIUS GEDIMINAS TECHNICAL UNIVERSITY

Ingrida MAŽONAVIČIŪTĖ

LITHUANIAN SPEECH ANIMATION TECHNOLOGY FOR 3D FACIAL MODEL

DOCTORAL DISSERTATION

TECHNOLOGICAL SCIENCES,
INFORMATICS ENGINEERING (07T)

VGTU
LEIDYKLA
Vilnius TECHNIKA 2012

Doctoral dissertation was prepared at Vilnius Gediminas Technical University in 2008–2012.

Scientific Supervisor

Prof Dr Habil Romualdas BAUŠYS (Vilnius Gediminas Technical University, Technological Sciences, Informatics Engineering – 07T).

VGTU leidyklos TECHNIKA d080-M mokslo literatūros knyga
http://leidykla.vgtu.lt

ISBN

© VGTU leidykla TECHNIKA, 2012
© Ingrida Mažonavičiūtė, 2012
ingrida.mazonaviciute@vgtu.lt

VILNIAUS GEDIMINO TECHNIKOS UNIVERSITETAS

Ingrida MAŽONAVIČIŪTĖ

LIETUVIŲ KALBOS ANIMAVIMO TECHNOLOGIJA TAIKANT TRIMATĮ VEIDO MODELĮ

DAKTARO DISERTACIJA

TECHNOLOGIJOS MOKSLAI,
INFORMATIKOS INŽINERIJA (07T)

VGTU
LEIDYKLA
TECHNIKA
Vilnius 2012

Disertacija rengta 2008–2012 metais Vilniaus Gedimino technikos universitete.

Mokslinis vadovas

prof. habil. dr. Romualdas BAUŠYS (Vilniaus Gedimino technikos universitetas, technologijos mokslai, informatikos inžinerija – 07T).

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. Accuracy of the speech recognition or generation engine, modelled visemes, phonemes to visemes mapping and coarticulation control model mainly influence speech animation quality, so phonetics of specific language must be analysed to create new „Talking head" (synthetic head model that can produce visual speech). Framework suitable to visualize Lithuanian language, which includes two new models to improve intelligibility of animated speech is proposed and applied to create Lithuanian „Talking head" „LIT".

The dissertation consists of Introduction, three main chapters and general conclusions.

Chapter 1 provides the analysis of the existing speech animation technologies. Different facial modelling techniques are examined to define the most suitable 3D „Talking head" modelling technique. Viseme classification experiments across different languages are analysed to identify variety of viseme identification and classification methods. Different coarticulation control models are compared to decide which one should be used to define coarticulation of Lithuanian speech.

Chapter 2 describes theoretical framework for Lithuanian speech animation. Translingual visual speech synthesis methods, specifics of Lithuanian phonetics and phonemes recognition system which is included in the proposed framework are analysed.

Chapter 3 presents the proposed Lithuanian speech animation technology and two new models to improve intelligibility of animated Lithuanian speech. The new translingual viseme selection model divides viseme classes of any language into two parts: constant (translingual) and variable. Technology to create constant (translingual) viseme set for 22 most popular consonants is proposed. Coarticulation control model specified for pure Lithuanian diphthongs is also proposed. The chapter is concluded with practical evaluation of the Lithuanian „Talking head" „LIT".

General conclusions summarise the present study. It is followed by an extensive list of 144 references and a list of 3 publications by the author on the topic of the dissertation.

Reziumė

Kalbos animacija naudojama technikos įrenginiuose norint sudaryti vienodas bendravimo galimybes kurtiesiems, vaikams, vidutinio ir vyresnio amžiaus žmonėms. Tai yra sudėtingas procesas, kurio metu žmogaus kalboje atpažinta akustinė informacija (fonemos) yra vizualizuojama naudojant specialiai sumodeliuotas veido išraiškas vadinamas vizemomis. Kuriant naujus kalbai animuoti skirtus modelius „Kalbančias galvas“ būtina atsižvelgti į analizuojamos kalbos fonetikos ypatumus. Kalbos animacijos tikroviškumui didžiausią įtaką turi tinkamas vizemų identifikavimas, modeliavimas ir pasirinktas koartikuliacijos valdymo modelis, naudojamas vizemoms laiko juostoje išrikiuoti. Karkasas, kuriame naudojama Lietuvių kalbai skirta animavimo technologija pasiūlytas ir realizuotas lietuviškai „Kalbančiai galvai“ „LIT“ sukurti.

Darbą sudaro įvadas, trys pagrindiniai skyriai, bendrosios išvados, literatūros ir publikacijų sąrašas.

Pirmame skyriuje analizuojamos esamos kalbos animavimo technologijos, tiriami skirtingose kalbose naudojami metodai vizemoms identifikuoti ir klasifikuoti, lyginami 4 pagrindiniai kalbos koartikuliacijos valdymo modeliai.

Antrajame skyriuje pristatomas karkasas, skirtas lietuvių kalbai animuoti. Išanalizavus esamus sprendimus, skyriuje siūlomas teorinis modelis, kaip atviro kodo anglų kalbos animavimo sistema gali būti panaudota lietuvių kalbos animavimo technologijai realizuoti.

Trečiame skyriuje detaliai aprašomi du nauji modeliai, kurie naudojami lietuvių kalbos animavimo technologijai sukurti ir lietuvių kalbos animacijos suprantamumui ir tikroviškumui padidinti. Tarpkalbinis vizemų parinkimo modelis yra naudojamas lietuvių kalbos vizemoms skirstyti į pastoviąją (tarptautinę) ir kintamąją dalis. Atsižvelgiant į lietuvių kalbos fonetikos taisykles pasiūlomas naujas koartikuliacijos valdymo modelis lietuvių kalbos dvibalsiams animuoti. Lietuviškai „Kalbančios galvos“ „LIT“ kokybė įvertinama praktiniais eksperimentais.

Darbas baigiamas išvadomis apie tyrimą. Disertacijos tema paskelbti trys straipsniai.

Notations

Abbrevations

2D – two dimensions;
3D – three dimensions;
FACS – Facial action system;
FAP – Facial animation parameter;
FAU – Facial action unit;
FDP – Facial definition parameter;
HMM – Hidden Markov Model;
HTK – Hidden Markov Model toolkit;
IPA – International Phonetic Alphabet;
MPEG-4 – Moving Picture Experts Group, 4 standard;
TTS – Text-To-Speech;
VCV – Vowel-Consonant-Vowel;
CVC – Consonant-Vowel-Consonant.

Contents

INTRODUCTION 1
Statement of the problem 1
The importance of the thesis 2
The objective of the thesis 3
The goal of the thesis 3
The tasks of the thesis 3
Research methodology 4
Scientific novelty 4
Practical significance of achieved results 4
The defended statements 5
Approval of the results 5
Dissertation structure 6
Acknowledgements 6

ANALYSIS OF CURRENT SPEECH ANIMATION TECHNOLOGIES 7
1.1. General terms 8
1.2. Audiovisual speech synthesis 8
1.2.1. Audiovisual speech generation methods 10
1.2.2. Speech processing for speech animation 17
1.2.3. Coarticulation modelling for audiovisual speech 21
1.2.4. Facial modelling for speech animation 24
1.3. Facial animation systems 28

1.3.1. Face parameterization and scripting languages 28
1.3.2. Current non-commercial facial animation systems..... 30
1.4. Facial animation evaluation methods..... 31
1.4.1. Advantages and disadvantages of evaluation methods 31
1.4.2. Challenges in evaluation of different speech animation systems..... 33
1.5. State of art in Lithuania 35
1.6. Conclusions on Chapter 1 38

FRAMEWORK FOR LITHUANIAN SPEECH ANIMATION 41
2.1. iFACE – speech animation system 42
2.1.1. Geometry of the head model..... 43
2.1.2. Behavior modelling and Face Modelling Language 45
2.2. Automatic phonemes recognition for the proposed framework..... 46
2.2.1. Specifics of Lithuanian phonetics..... 46
2.2.2. Framework for Lithuanian speech recognition module 47
2.3. Translingual visual speech synthesis 53
2.4. Proposed architecture to animate Lithuanian speech 55
2.5. Conclusions on Chapter 2 58

PROPOSED LITHUANIAN SPEECH ANIMATION TECHNOLOGY..... 61
3.1. Translingual visemes selection model 62
3.1.1. Determination of constant part of viseme classes..... 62
3.1.2. Analysis of variable part of viseme classes 74
3.1.3. Application of translingual visemes selection model 82
3.1.4. Concluding remarks on translingual visemes selection model 84
3.2. Lithuanian coarticulation model for pure Lithuanian diphthongs..... 84
3.2.1. Coarticulation analysis of Lithuanian diphthongs 84
3.2.2. Proposed framework for pure Lithuanian diphthong visualization..... 86
3.2.3. Concluding remarks on coarticulation modelling for pure Lithuanian diphthongs 89
3.3. Practical evaluation of the proposed framework..... 90
3.3.1. Environment of the experiment 90
3.3.2. Intelligibility testing of animated speech..... 91
3.3.3. Naturalness testing of animated speech 93
3.3.4. Concluding remarks on practical evaluation of the proposed framework 96

GENERAL CONCLUSIONS 97

REFERENCES 99

LIST OF PUBLICATIONS BY THE AUTHOR ON THE TOPIC OF THE DISSERTATION 107

List of figures

Fig. 1.1. General pipeline of data-driven speech animation. The sample-based approaches go in the top path and the learning based approaches in the bottom path. ... 12
Fig. 1.2. Viseme-driven approach for visual speech animation, when input for speech processing module is recorded speech. ... 13
Fig. 1.3. Focus points used to identify 4 geometric parameters that were employed to define visemes for Standard Malay language. ... 15
Fig. 1.4. Geometric parameters used to didentify visemes for Modern, Classical Arabic language. W depicts inner lips width, H – inner lips height, A – inner lips area and D – the area of dark zone. ... 15
Fig. 1.5. Framework of the typical Text-to-Speech system ... 17
Fig. 1.6. Recognition of natural and synthetic „Talking head" joined with natural and synthetic speech ... 18
Fig. 1.7. The common structure of the Automatic Speech Recognition system ... 20
Fig. 1.8. The dominance of three concatenated speech segments over time (top) and the resulting synthesized trajectory of one parameter (bottom) ... 22
Fig. 1.9. "The uncanny valley" effect. Hypothesized emotional response of human subjects is plotted against anthropomorphism of a synthetic face. The uncanny valley is the region of negative emotional response towards syntetic faces that seem "almost human" ... 34
Fig. 1.10. Selection scheme for „Talking head" modelling technique that is the most suitable for Lithuanian „Talking head", which is based on translingual adaptation. Black rectangular mean the best choice in that level ... 39

Fig. 2.1. Architecture of iFACE speech animation system, originaly created for English language 43
Fig. 2.2. Hierarchy of abstraction layers that defines geometry of the head model used in iFACE speech animation system 44
Fig. 2.3. Parameters spaces that determine communicative behaviour of the face in iFACE system 45
Fig. 2.4. Continuous speech recognition framework for Lithuanian speech recognition 48
Fig. 2.5. Phonetic transcription of the word "akti" from the Praat annotation file 51
Fig. 2.6. Main window of Praat. Speech waveform and spectrogram (with a pitch contour) of the same sentence with annotations on phoneme-level and word-level are shown 52
Fig. 2.7. Block diagram showing the modification layers of translingual speech synthesis 55
Fig. 2.8. Architecture of the proposed Lithuanian speech animation framework 56
Fig. 3.1. Interlanguage confusion matrix for consonant, where darker zones means higher visual correlation between phonemes in the ith row and the jth column 66
Fig. 3.2. Dendragram that illustrates consonants clustering into viseme classes with reference to visual similarity of phonemes, that were defined in viseme classification tables in 11 languages. Group-average agglomerative clustering method was used to define similarity between two clusters 69
Fig. 3.4. Similarity between new interlanguage consonants-to-viseme mapping table and phoneme-to-viseme mappings for different languages 73
Fig. 3.5. Visemes classification in „Crazy Talk" software 74
Fig. 3.6. Image acquisition layout for the experiment 76
Fig. 3.7. 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 79
Fig. 3.8. 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 80
Fig. 3.9. 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. 81
Fig. 3.10. Framework for pure Lithuanian diphthong animation. Dotted and greyed rules define situation when Ohman coarticulation control model can be applied and white rules define state, where it cannot be done 86
Fig. 3.11. Expressiveness parameter before and after application of our proposed model 89
Fig. 3.12. The percentage of correct words for the different subjects in the three modalities (unimodal auditory, audiovisual speech using „iFACE" and „LIT" speech animation systems) 92

List of tables

Table 1.1. Summary of intelligibility test of visual speech synthesis coarticulation control models (Bescow, 2004) 24
Table 1.2. Comparison of Video-based and 3D models based facial animations in terms of automastism, flexibility and realism. 27
Table 1.3. Content description languages and their features 30
Table 2.1. Major LRN0 characteristics 49
Table 2.2. Phoneme error rates (PER) for the simplest Lithuanian phoneme set 50
Table 3.1. Consonants and their matching viseme groups in 11 languages. 64
Table 3.2. Lithuanian viseme to English viseme mapping table. 82
Table 3.3. Average Lithuanian words recognition rate in in 3 distinct conditions 93
Table 3.4. An average naturalness result on videos generated before and after English speech animation system was adapted to animate Lithuanian words. Likert skale, where very natural visual speech is scored as 5, was used for the estimation..... 96

Introduction

Statement of the problem

Human communicate using words and sentences. Lipreading is considered as a psychological process not only in people with hearing loss and deafness, but also in normal hearing people. Visual information (i.e. lips and tongue movements) significantly enhances the comprehension of the uttered audio signal. This is particularly true when the acoustic conditions are inadequate. Researchers revealed that addition of the visual information increases the speech perspicuity with 57% for consonants, 30% for vowels, 39% for monosyllabic words and 17% for short phrases.

Ever since visual speech improves the perception of acoustical signal, design and implementation of 2D or 3D synthetic head models that can produce naturally looking audio to video mapping ("Talking heads") became very important in Human – Computer Interaction area. Design of "Talking heads" indistinguishable from a real human is challenging task, since people are highly sensitive to facial behaviour. Moreover, casual viewer immediately perceives the slightest deviation from the natural facial appearance and defines it as inaccurate visualization.

During the last years, different „Talking heads" have been proposed, pursuing either a natural 3D facial mesh or video-realistic effects. Most of the existing

models are dedicated to animate English language, although recently we also see "Talking heads" driven by Finnish, Italian, Chinese Mandarin (Putonghua) and etc. However, there is not „Talking head" specified to visualize Lithuanian language.

The importance of the thesis

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. "Talking heads" 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, internet navigator or virtual agent who is responsible for information conveying to user in a Smart Ecological and Social Apartments (SESA). Also they are widely used in e-learning technologies as animated 3D models for the correct sound pronunciation presentation or applied in movie, advertising and computer game industries.

"Talking heads" are basically driven by speech phonetics and their visual representation – visemes. Many acoustic sounds of separate languages are visually similar and accordingly different phonemes can be classified using the same viseme. Phoneme to viseme mapping remains the key subject in bimodal speech processing due to the fact that correctness of identified visemes considerably influence the realism of animated speech. Despite the fact, that various researchers attempted to classify phonemes according to their visual similarity, as of date, neither a single classification, nor a standardized viseme set exist for individual languages, as it exists for phonemes. For this reason audio to visual mapping should be revised for every new language, even if it is expensive and time consuming process.

In 2005, Lithuanian scientist Balbonas proposed classification of Lithuanian visemes by their phonemic distinction into explosive, flint, guttural, etc. In this approach, differently apparent Lithuanian consonants /d, t, s, z, n, dz, l/ belong to the same viseme class. It causes speech animation artificiality, thus additional research should be done to define Lithuanian viseme classes.

Since all humans are experts in lip reading and detects even the slightest errors during speech animation, coarticulation generation is crucial part of any speech animation system. Coarticulation is the physical phenomena which refers to the blurring of boundaries between separate speech units (both auditory and visibly). Word stressing, position of the accented phoneme, pronunciation of diphthongs, long and short vowels highly influence final visualization of the

speech. Lithuanian coarticulation control model, which is based on Lithuanian phonetics, is proposed to animate words with pure diphthongs.

Proper visemes definition and their correct alignment are the crucial points for believable language animation. The proposed coarticulation control and translingual visemes selection models are used to create Lithuanian „Talking head" „LIT".

The objective of the thesis

The objective of the thesis is Lithuanian speech animation.

The goal of the thesis

The main goal of the thesis is to create Lithuanian speech animation technology

The tasks of the thesis

In order to achieve the goal, the following tasks had to be solved:

1. To define speech animation method and „Talking head" modelling technique that is the most perspective for Lithuanian speech animation. To indicate open source speech animation system that has these features.
2. To create Lithuanian speech animation platform, which preprocesses, synchronizes and integrates acoustic and visual features of Lithuanian speech.
3. To propose technology, that designates how to divide viseme set of any language into two parts: constant and variable. Translingual viseme set for globally the most common consonants should be determined to identify steady set of visemes.
4. To create Lithuanian coarticulation control model that concerns phonetic features of Lithuanian language and which is specified to animate pure Lithuanian diphthongs.
5. To evaluate the effectiveness of the proposed speech animation technology considering intelligibility and naturalness of created Lithuanian „Talking head" „LIT".

Research methodology

Theoretical analysis and practical implementation knowledge of mathematical analysis, speech animation methodologies, phoneme to viseme mapping techniques, coarticulation generation and 3D modelling by calibration from 2D pictures were used. The base software for practical implementation of the experiments was English speech animation system „iFACE" and its Facial Modelling Language (FML).

Scientific novelty

The thesis addresses the problem how viseme-driven parametric „Talking head" can be employed to animate new languages. Lithuanian language is analysed in this thesis:

1. A novel Lithuanian speech animation technology, that employs parameter-based, 3-dimensional, speech driven Lithuanian „Talking head" „LIT", was created. It improves intelligibility and naturalness of animated Lithuanian speech.
2. The new speech animation technology was created, that allows division of viseme set of any language into constant (translingual) and variable parts. It considerably reduces the amount of visemes that should be analysed or modelled for specific language. Translingual classification of visemes for 22 internationally common consonants can be applied to produce language independent "Talking heads".
3. New coarticulation control model, that defines eight rules for pure Lithuanian diphthong animation, was created. Original coarticulation generation methodology to animate structures VVCV (vowel-vowel-consonant-vowel) was proposed and incorporated into the model.

Practical significance of achieved results

The proposed Lithuanian speech animation technology with integrated translingual phoneme to viseme mapping module can be successfully applied to model new language-independent viseme-driven "Talking heads". Lithuanian speech animation platform can be used to create Lithuanian off-line „talking head" that can be employed in e-learning technologies as animated 3D model for the correct sound pronunciation presentation or applied in movie, advertising and computer game industries. The presented Lithuanian coarticulation model-

ling technique can be used to increase naturalness of animated speech for other languages with many diphthongs.

The defended statements

The following defendable statements are defined:

1. The proposed speech animation platform is suitable to visualise recorded Lithuanian speech.
2. Translingual viseme classification technology, that allows division of viseme set of any language into constant (translingual) and variable parts, can be employed to create new viseme-driven „Talking heads".
3. The proposed Lithuanian coarticulation control model is suitable to animate Lithuanian words with pure diphthongs.

Approval of the results

There are 3 scientific articles posted that are published based on results of this thesis: 2 Articles in ISI Web of Science (Mazonaviciute et al, 2009, Mazonaviciute et al., 2011., 1 Conference Paper in Reviewed International Conference (Mazonaviciute et al., 2010).

The author has made 3 presentations at 3 international scientific conferences in Lithuania and abroad:

18th European Signal Processing Conference (EUSIPCO-2010), Aalborg, Denmark, August 23–27, 2010.

15th International conference of Electronics (Electronics'2011), Kaunas and Vilnius, Lithuania, May 17–19, 2011

16th International conference of Electronics (Electronics'2012), Palanga, Lithuania, June 18–20, 2012

II Award in the contest „The Best presentation from the young scientist". For the presentation „Translingual visemes mapping for Lithuanian speech animation". IEEE, 2011.

Dissertation structure

The dissertation consists of Introduction, three main chapters and general conclusions. The total dissertation scope is 107 pages, 10 equations, 30 pictures and 9 tables.

General conclusions summarises the present study. It is followed by an extensive list of 144 references and a list of 3 publications by the author on the topic of the dissertation.

Acknowledgements

I am grateful to my scientific supervisor Prof Dr Habil Romualdas Baušys for guiding and helping me in my studies and preparing the thesis. I would also like to thank for all the academic stuff of Graphical systems cathedral for support and valuable comments. This thesis would not have been possible unless support from my mother Genovaite Mazonaviciene and my friends.

1

Analysis of current speech animation technologies

In this chapter a review of current state-of-art in audiovisual speech generation techniques and facial modelling, suitable for speech animation, is presented. Coarticulation modelling and acoustic to visual mapping is reviewed broader, since their accurateness highly influences the naturalness and intelligibility of animated speech. One of the main tasks of this thesis is to enlighten new aspects of translingual phoneme to viseme mapping; therefore special attention is paid to identification methods of viseme classes in different languages. The second important task of our research is to propose speech visualization framework that employs translingual speech animation techniques. For this purpose survey on current non-commercial facial animation systems with face parameterization and scripting languages is carried out.

Speech processing module is very important part of any speech animation system, so it is briefly reviewed in this state-of-art. Evaluation methods of "Talking heads" are presented also, because estimation of animated speech greatly depends on the human observer.

Chapter is started with definition of general terms and concluded with state-of-art review in Lithuania.

1.1. General terms

Speech animation. Speech is produced by the movements of specific organs or articulators of the vocal tract. Each distinct speech sound is related to characteristic positions of the articulators, and some of their movements are wholly or partially visible on the speaker's face, especially in the region around the mouth, which comprises the upper end of the vocal tract. Speech animation is the simulation of such visible movements during speech production. This term also refers to techniques for specifying and controlling the positioning and movement of a synthetic face into and between facial expressions. The portrayal of all other possible facial behaviours, such as emotions and speech related communication signs, that are not strictly bounded to the physical restrictions of speech production are the large part of speech animation (Pelachaud et al., 1996).
Phoneme. Phoneme is a family of similar sounds that a language treats as being "the same". Also it is understandable as a minimal acoustic unit, which serves to distinguish between meanings of words. Diphones are two consecutive phonemes and triphones are three consecutive phonemes. A complex speech sound or glide that begins with one vowel and gradually changes to another vowel in the same syllable (diphthong) is also defined as phoneme.
Viseme. Human eye cannot visually differentiate between pronunciation of different sounds like /p/ and /b/, while both of them are plosives and look analogous during the speech. Consequently, visually analogous phonemes can be compound into a single viseme class that is illustrated by specific lip shape. Usage of viseme classes reduces the amount of facial poses that must be defined to produce video-realistic speech animation.
Speech coarticulation. The assimilation of the place of articulation of one speech sound to that of an adjacent speech sound. For example, while the sound /n/ of English normally has an alveolar place of articulation, in the word tenth it is pronounced with a dental place of articulation because the following sound, /θ/, is dental. Usually coarticulation is understandable as the differing of phonemes depending on context.

1.2. Audiovisual speech synthesis

The human face is a very complex surface and extremely important communication channel. The face can express lots of information, such as emotions, intension or general condition of the person. There are three types of facial cues carrying out different communication tasks. First group expresses emotions such as sadness, anger, joy, fear, disgust and surprise. Second one occurs in verbal communication situations and expresses information about discourse, phrasing,

emphasis and dialogue turn-taking, which are often related to intonation features of voice. Finally, the third group of facial cues accompany acoustic speech as visible speech and carry information about phonetic content. Since the last category contains the visual information about speech, it significantly improves speech animation intelligibility and plays a key role for hearing impaired people (Campbell et al., 1987). However, lip-reading is considered as a psychological process not only in people with hearing loss and deafness, but also in normal hearing people. Lip movements can compensate a possible loss in speech signal especially in noisy environments.

Use of multiple sources (acoustic and visual) generally enhances a speech perception and understanding. The McGurk (McGurk et al., 1976) demonstrated strong relation between acoustic and visual signal for speech perception. For example, listening of syllable /ba/ and observing the pronunciation of syllable /ga/ at the same time may be psychologically perceived as /ga/ or /da/. Such perceptual results are repeatable among subjects across age, gender and native language. So, visual information (i.e. lips and tongue movements) significantly enhances the comprehension of the uttered audio signal. Researchers revealed that addition of the visual information increases the speech perspicuity with 57% for consonants, 30% for vowels, 39% for monosyllabic words and 17% for short phrases (O´Neill, 1954).

People speak using articulators like jaw, lips, tongue, velum and larynx, nasal cavity and oral cavity. Visual speech is produced as the continuous flow of articulatory movements, written as a discrete set of phonetic symbols. These symbols include:

a) Phonemes – the minimal units of speech that serve to distinguish between meanings of words. Diphones are two consecutive phonemes while triphones are three consecutive phonemes

b) Allophones – the members of the phoneme family. Allophone is a phonetic variant of a phoneme in a particular language.

c) Monophthongs – a single uncompounded vowel sound or combination of two written vowels pronounced as one.

d) Diphthongs – a complex speech sound or glide that begins with one vowel and gradually changes to another vowel or consonant in the same syllable.

Basic facial animation can be considered as the concatenation of the visual equivalent of speech segments like phonemes, otherwise known as visemes. There is no simple one-to-one relation between phonemes and visemes, since different sounds may look the same. An example of this case is two bilabial stops /b/ and /p/. Visually there is no difference in fluent speech between them,

so these two phonemes can be placed in the same visemic class. However there is no standard set of viseme classes and the count of visemes and their content changes in different investigations.

Coarticulation phenomena also complicate mappings between acoustic speech signals and visual speech motions and it is highly related with phonemic and visemic vocabularies of different languages.

1.2.1. Audiovisual speech generation methods

Previous research efforts in audiovisual speech generation (speech animation) can be roughly classified into two different categories (Deng et al., 2007): data-driven approaches and viseme-driven approaches.

Data-Driven facial animation

Data-driven approaches synthesize new speech animations by concatenating pre-recorded facial motion data or sampling from statistical models learned from the data. 3 main phases should be completed to generate data-driven speech animation:

1. Recording of facial motions. It can be 2D facial images or 3D facial motion capture data.
2. Processing of recorded facial motion. There are two different ways to deal with the constructed facial motion database. Either facial motion database is further organized and processed (*sample-based approaches* (Bregler et al., 1997), (Cosatto, 2003), (Kshirsagar et al., 2003)), or statistical models for facial motion control are trained from the recorded data (*learning-based approaches* (Ezzat et al., 2002), (Meyer, 2003), (Deng et al., 2006)).
3. Realization of speech synthesis. Recombination of motion frames optimally chosen from the facial motion database or sampling from the trained statistical models is used to animate novel sound track or text input.

Sample-based approaches. Sample based facial animation concatenates phoneme or syllable segments from a pre-recorded facial motion database. Bregler et al. (1997) present the “video rewrite“ method for synthesizing 2D talking faces given novel audio track, based on a collected database of “triphone video segments“. Instead of using ad hoc co-articulation models and ignoring dynamics factors in speech, this approach models the coarticulation effect with “triphone video segments”, but it is not generative (i.e. the co-articulation cannot be applied to other faces without retraining). The success of this approach largely

depends on the accuracy of phoneme-alignment and the number of triphones covered in the training video footage. Cao et al. (2004) proposed a greedy search algorithm to look for longer pre-recorded facial motion sequences (longer than 3 phonemes) in the database.

Instead of constructing a phoneme segment database, Kshirsagar et al. (2003) present a syllable based approach to synthesize novel speech animation. In their approach, captured facial motions are chopped and categorized into syllable motions, and then new speech animations are generated by concatenating corresponding syllable motions from the created syllable motion database.

Sifakis et al. (2006) proposes a physics-based approach to generate novel speech animations by first computing muscle activation signals for each phoneme (termed as physemes) enclosed in the pre-recorded facial motion data and then concatenating corresponding physemes given novel speech input.

Learning-based approaches. Learning based approaches learn statistical models from data for facial animation synthesis and editing. Brand (1999) learns HMM-based facial control models by an entropy minimization learning algorithm from voice and video training data and then effectively synthesizes full facial motions from novel audio track. This approach models co-articulations, using the Viterbi algorithm through vocal HMMs to search for most likely facial state sequence that is used for predicting facial configuration sequences .Ezzat et al. (2002) learn a multidimensional morphable model from a recorded video database that requires a limited set of mouth image prototypes and use the magnitude of diagonal covariance matrices of phoneme clusters to represent co-articulation effects: the larger covariance of a phoneme cluster means this phoneme has a smaller coarticulation, and vice versa.

Chuang et al. (2005) learn a facial expression mapping/transformation from training footage using bilinear models, and then this learned mapping is used to transform novel video of neutral talking to expressive talking. Cao et al. (2003) present a motion editing technique that applies Independent Component Analysis (ICA) to recorded facial motion capture data and further editing and mapping operations are done on these ICA components, interpreted as expression and speech components separately. Zhang et al. (2003) present a geometry-driven technique for synthesizing expression details for 2D faces, where users can move a small number of control points on 2D face images, and then movements of other 2D control points are automatically computed using a motion propagation algorithm.

The schematic view of both data-driven speech animation approaches is demonstrated in Fig. 1.1.

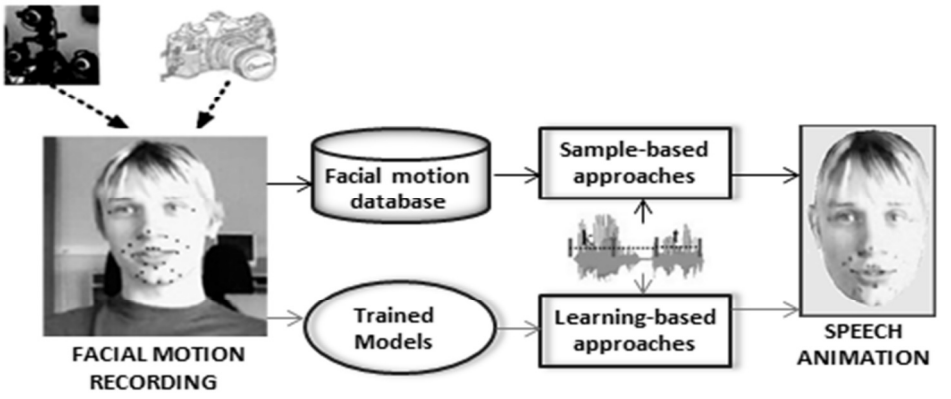


Fig. 1.1. General pipeline of data-driven speech animation. The sample-based approaches go in the top path and the learning based approaches in the bottom path.

Data-driven approaches typically generate 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 animators. Model and feature selections residing in many machine learning algorithms are also a challenge for the researchers.

Viseme-Driven speech animation

Viseme-driven speech animation is based on the fact, that there are many sounds that are visually ambiguous when pronounced. Using this approach generation of novel speech animations consists of 3 main steps:

1. Speech recognition or text-to-speech engine is used to recognize phonemes sequence and to detect their timing information.
2. Audio to visual mapping vocabulary is created. Various methods can be used to identify visemes classes and to relate them with corresponding phonemes.
3. Phoneme to viseme mapping is used to visualize speech. Visemes are synchronized with phonemes using empirical smooth functions (Cosi et al., 2002; Roach, 2004) or coarticulation rules (Bescow, 1995).

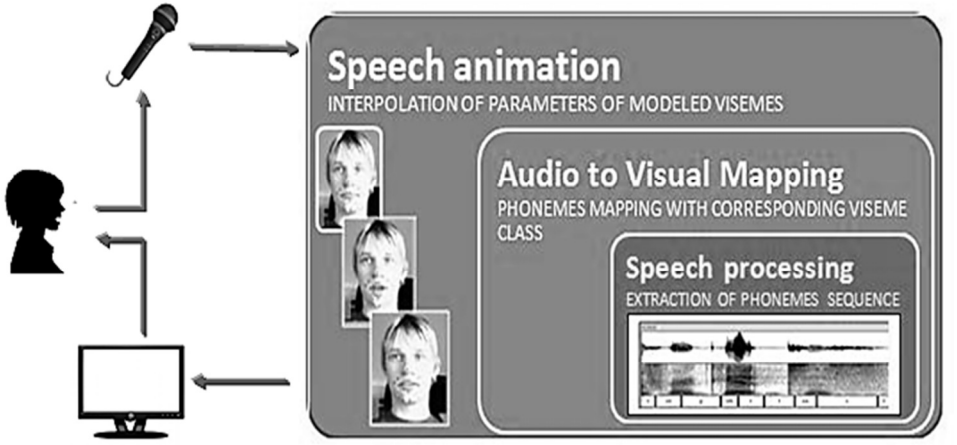


Fig. 1.2. Viseme-driven approach for visual speech animation, when input for speech processing module is recorded speech.

The base idea of viseme-driven approach is demonstrated in Fig. 1.2. In this approach final animation depends on three main subjects: precision of speech processing engine, correctness of phonemes to visemes mapping and the choice of coarticulation model. Audio to visual mapping module is the key feature in viseme-driven animation. General methods for speech processing or coarticulation modelling, that can be applied both to data-driven and viseme driven animation will be analysed as separate subsections of Chapter 1.

Viseme classification in different languages. As of date, neither a single classification method for visemes, nor a standardized viseme set exist for individual languages, as it exists for phonemes (Damien et al., 2009). Variability on the number of English viseme classes confirms this assertion: Massaro described 17 viseme groups (Massaro, 1998), meanwhile Dodd and Cambell (Dodd et al., 1984) as well as Tekalp and Ostermann (Tekalp et al., 2000) identified 14 viseme classes.

Visual phonemes classification into viseme classes is a vast problem since the selection of appropriate video segments is based on visemically transcribed phonemes. Approaches to identify viseme classes can be roughly classified into two general groups (DeMartino et al., 2006):

1. Approaches focused on perceptual issues.
2. Approaches concerned with measurable geometric aspects visible on the face.

First category cover early observations derived from clinical experience with hearing impaired individuals and various studies based on visual recognition tests (Owens et al., 1985). In 1968, Fischer (1968) made a pioneer research on visemes identification and tested how English phonemes were perceived visually depending on their position in the word. For this experiment he defined additional words that had the same syllabic structure and stress pattern as the examined word. Finally, the correct answer was removed from the answer sheet, so participants were forced to give wrong responses during the process of word recognition. Investigation of the answer sheets confirmed that phonemes were confused for ten groups of consonants. These groups support the concept of viseme classes.

The second category is based on measurements of movements displayed on the face during speech production. These approaches are based on the understanding that viseme classes involve the grouping of phonemes according to similarities in visible geometry. A number of techniques have been used to capture and model key mouth shapes for speech animation. Parke et al. (1996) employed rotoscoping of recorded frames of a real speaker to define speech poses for demonstration utterances. Hill et al. (1988) derived viseme parameters from static photographs illustrating a book about lip-reading. Waters and Levergood (1993) defined a chart of mouth shapes constructed from the observation of real lips.

Different features can be selected to explore geometrical transformation of lips while man is talking and to identify separate classes of visually similar phonemes. The choice of these features is a crucial, since it's the primary step of the visemes identification process. Different features were used to identify visemes across separate languages.

Abry and Boë (1986) proposed a set of eight geometric parameters to describe lip shapes. Three of them were defined in front view of captured speaker and five of them – in profile view. The same count of parameters was chosen to identify French viseme set (Le Goff et al., 1996).

Automatic lip reading system proposed by Chen (2001) was employed to define phoneme-to-viseme mapping table for Standard Malay language (Salleh et al., 2010). The experiment, when native speaker pronounces earlier defined phrases was recorded to collect data for this investigation. Lip movements that were used as visual characteristics were combined with audio features to execute recognition of visual phonemes using hidden Markov model (HMM). Audio signal was divided into smaller frames and converted into sixteenth-order linear prediction coding (LPC) coefficients to create audio features; in the meantime colour distribution of the face pixels was observed for the lip-tracking phase. After all, the template resembling the shape of the lips was applied to extract 4 geometric parameters and exploit them for definition of 14 viseme classes. 4

feature points that were used to identify two lip corners and height of the upper and lower lips are illustrated in Fig. 1.3.

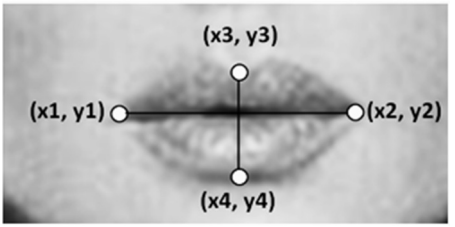


Fig. 1.3. Focus points used to identify 4 geometric parameters that were employed to define visemes for Standard Malay language.

European Portuguese viseme classes (Neto et al., 2006) were defined after examination of 3 physical natural descriptors that are lip height, lip width and distance between chin point and top lip. Meanwhile correspondence between phonemes and visemes in Modern, Classical Arabic language (Damien et al., 2009) is based on 4 geometric parameters: three front features (width (W), height (H) and area of inner lips (A)) and additional parameter corresponding to the area of dark zone (D). The additional feature helps to recognize phonemes whose pronunciation requires a contact between the tongue and the teeth. Parameters W, H, A and D are demonstrated in Fig. 1.4.

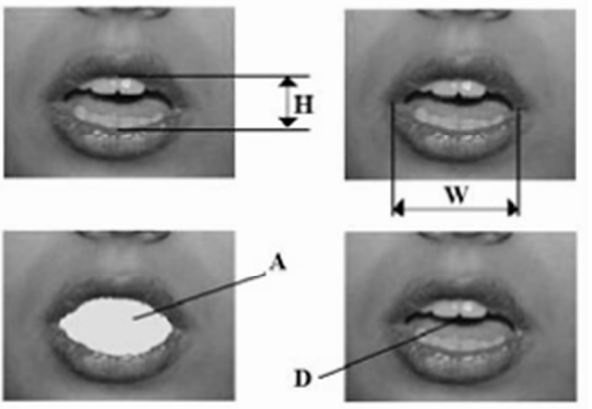


Fig. 1.4. Geometric parameters used to didentify visemes for Modern, Classical Arabic language. W depicts inner lips width, H – inner lips height, A – inner lips area and D – the area of dark zone.

In Italian (Magno-Caldognetto, 1998), 5 features points (central points of upper and lower lips together with markers placed on jaw and lobes of the ears) were analysed to calculate 3 important parameters which are used for visemes identification. Those parameters are lip height, lip width and lip protrusion. Lip protrusion is calculated as the distance between the marker placed on the central points of either upper and lower lip and the frontal plane containing line crossing the markers placed on the lobes of the ears. Values of articulatory parameters may be either positive or negative: for instance, lip weight values are negative when the distance between the corners of the lips decreases with respect to their distance at rest (lips are rounded). Based on these values statistical definition of viseme classes for Italian vowels and consonants were performed. Finally, 7 Italian vowels were grouped into 4 viseme classes and 21 consonants were divided into 6 groups of visually similar phonemes.

However, a number of investigators extracted viseme classes without clear definition of feature points. For example, German viseme identification process (Aschenberner et al., 2005) is based on subjective observation of the places of articulation during the pronunciation of phonemes. The place of articulation is the point where an obstruction occurs in the vocal tract between an articulatory gesture, an active articulator and a passive location. The passive location is the place on the more stationary part of the vocal tract where the articulation occurs. It can be anywhere from the lips, upper teeth, gums, or roof of the mouth to the back of the throat. The articulatory gesture of the active place of articulation involves the more mobile part of the vocal tract. This is typically some part of the tongue or lips. German viseme classes relate to articulators that are visibly contrastive. For instance, two bilabial plosives /p, b/ are defined as single viseme and the four remaining plosives /t,d,k,g/ are put together as different viseme class. In such a manner 42 German phonemes are mapped to 15 viseme classes.

For the research of Mexican-Spanish (Lazalde et al., 2007) viseme classes recording of experiments and data processing were not performed at all. Definition of 15 viseme classes for 30 Mexican-Spanish phonemes is based on comparative analysis of differences between phonetics of English and Mexican-Spanish languages. For example, in Spanish, the phoneme /v/ is mapped with phonemes /b, p, m/, in the meantime /v/ and /f/ are defined as visually similar in English. Some Spanish phonemes like /ñ/ do not occur in English at all. In such a case, the appropriate viseme class in English phoneme-to-viseme mapping table is found to describe this Mexican-Spanish phoneme. Finally, letter /h/ it is not pronounced in Spanish, so it doesn't have a corresponding phoneme and viseme.

Chinese Mandarin is the world's most widely spoken language, which has a distinctive phonetic system. Chinese Mandarin (Hoon, 2004) is a monosyllabic dialect, where each syllable consists of an optional initial (i.e. a consonant), final

(i.e. a core vowel followed by an optional consonant) and a tone. Furthermore, finals are classified into single finals and complex finals. So, there are 21 initials and 38 finals in Chinese Mandarin, where every initial as well as single final can be represented by a static viseme. 20 viseme classes for initials and finals were defined in (Wu et al., 2006). Complex finals are formed by a combination of corresponding static visemes. Tone is not analysed since it doesn't affect visemes. The Hanyu Pinyin system (Wu et al., 2006) is used to represent each Chinese character by a string of Latin letters.

Viseme-driven speech animation approach requires challenging and careful manual work to identify phoneme-to-viseme mappings; however, intuitive management of visual speech synthesis and possibility to re-use viseme models for a new "Talking heads" is very important for translingual speech animation.

1.2.2. Speech processing for speech animation

„Talking heads" can be driven by natural or synthetic speech. Text-to-speech systems (TTS) can be used to convert written text into synthetic speech. A TTS system is generally composed of two parts (Van Santen et al., 1997): a front-end and a back-end.

The front-end has two major tasks. First, it converts raw text containing symbols like numbers and abbreviations into the equivalent of written-out words. The front-end then assigns phonetic transcriptions to each word, and divides and marks the text into prosodic units, like phrases, clauses, and sentences.. Phonetic transcriptions and prosody information together make up the symbolic linguistic representation that is output by the front-end.

The back-end converts the symbolic linguistic representation into sound. In certain systems, this part includes the computation of the target prosody (pitch contour, phoneme durations), which is then imposed on the output speech.

Schematic view of a typical TTS system is presented in Fig. 1.5.

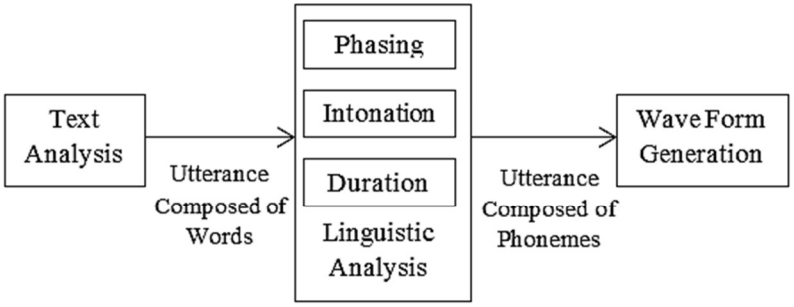


Fig. 1.5. Framework of the typical Text-to-Speech system

Analysis of natural and synthetic "Talking heads" joined with natural and synthetic speech (Bescow, 1995) proved that recognition of natural speech combined with synthetic face is higher than recognition of natural speech alone (Fig. 1.6.). However, it also proves that intelligibility of synthetic face combined with synthetic speech is 25% worse than perspicuity of synthetic face combined with synthetic speech. This is the reason why natural speech should be the input for Lithuanian „Talking head" called "LIT".

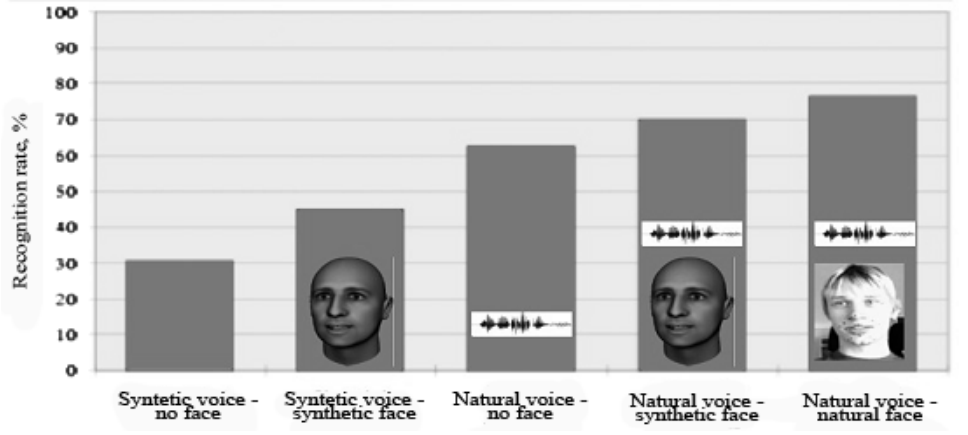


Fig. 1.6. Recognition of natural and synthetic „Talking head" joined with natural and synthetic speech

The problem of mapping from the audio feature space to the visual feature space can be solved at 3 different levels (Huang et al., 1998):

a) Signal level (front end);

b) Phoneme level (acoustic model);

c) Word level (language model).

AV mapping at signal level concentrates on the physical relationship between the shape of the vocal tract and the produced sound. Speech signal is segmented into frames and AV mapping is performed frame by frame. There are many algorithms that can be modified to perform this kind of mapping – Vector Quantization (VQ), the Neural Network (NN) (Massaro et al., 1999), the Hidden Markov Model (HMM) (Brand, 1999), etc. Signal level mapping techniques are incorporated in AV mapping for each higher hierarchical state.

At the phoneme level, the mapping is done for each phoneme in the speech signal. The first step of this technique is to segment the speech sequence phonet-

ically. Then we use a lookup-table to find out the sequence of visual features. Usually phoneme-to-viseme mapping table, which contains individual visual feature set for each phoneme, is used to define AV mapping at phoneme level. This kind of AV mapping extremely depends on the quality of phonemes recognition engine.

At the third – word level, the context cues in the speech signals can be explored. The speech recognizer is used to segment speech into words. Hidden Markov Model (HMM) (Young et al., 2000) should be created to represent acoustic state transition in every word. For each state in the HMM model, signal level mapping methods can be applied.

The choice of the particular level depends on application where AV mapping should be used. Signal level is simple, language independent and suitable for real-time implementation, but contrary to the latter two approaches, coarticulation are not incorporated. AV mapping at phoneme and word levels are context depended and provides more precise speech analysis. Word level is tailored to individual words, so it can produce better results than signal-level or phoneme-level approaches. However, since input signal of higher level requires more complex speech recognition system, phoneme level is faster and simpler then word level.

Efficiency of AV mapping at phoneme level exceptionally depends on the chosen speech recognition engine, usually called Automatic Speech Recognition system (ASR). ASR performs two fundamental operations: signal modelling and pattern matching (Picone, 1993).

Signal modelling represents process of converting speech signal into a set of parameters and involves four basic operations:

a) Spectral shaping – the process of converting the speech signal from sound pressure wave to a digital signal and emphasizing important frequency components in the signal.
b) Feature extraction – the process of obtaining different features such as power, pitch, and vocal tract configuration from the speech signal.
c) Parameter transformation – the process of converting extracted features into signal parameters.
d) Statistical modelling that involves conversion of parameters into signal observation vectors.

Pattern matching is the task of finding parameter set from memory which closely matches the parameter set obtained from the input speech signal. It can be done in word or sentence level. Acoustic model of words together with Language model are used for that.

Acoustic models of words are the result of acoustic modelling – a process of building and development of acoustic models according to the sub-word units derived from a linguistic criteria and training data (Laurinciukaite, 2009). Each

set of acoustic model has no less than one lexicon, which includes all the words used in the ASR system modelling and subsequent recognition task. The lexicon gives the transcription of a word in a meaningful sequence of sub-word units (each sub-word unit has an acoustic model).

Language models compare syntax and semantics of the analysed language. The common structure of the ASR system is shown in Fig. 1.7. (Laurinciukaite, 2009).

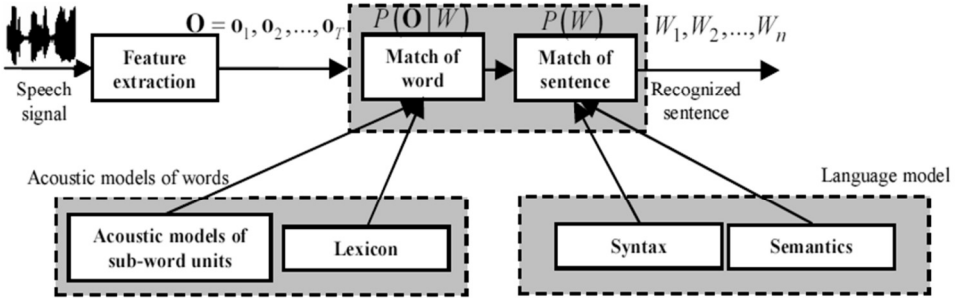


Fig. 1.7. The common structure of the Automatic Speech Recognition system

The dominant speech recognition paradigm of recent years is known as Hidden Markov Models (HMM) (Young et al., 2000). HMM is a statistical Markov model in which the analysed system is assumed to be a Markov process with hidden states. Although the state is not directly visible to the observer, the output, dependent on the state, is visible. Each state has a probability distribution over the possible output tokens. HMM-based speech recognition engine identifies speech by estimating the probability of each phoneme at contiguous small frames of the speech signal. A search procedure is used to determine the sequence of phonemes with the highest probability. This search is constrained to look for phoneme sequences that correspond to words in the vocabulary and the phoneme sequences with the highest total probability are identified with the spoken word.

The described Automatic speech recognition methodology can be applied to process speech in different languages; however each language has its own unique features. It results that individual speech recognitions engines must be created for distinct languages.

1.2.3. Coarticulation modelling for audiovisual speech

Viseme-driven animation depends on three main subjects: precision of speech processing engine, correctness of phonemes to visemes mapping and the choise of coarticulation model. Coarticulation is very important for data-driven animation also.

Coarticulation is the physical phenomena which refers to the blurring of boundaries between separate speech units (both auditory and visibly). For instance, when the nasal consonant [m] precedes an oral vowel [a] in the English word "map", some of the nasality will carry forward on the vowel [a]. During the pronunciation of the nasal [m] the soft palate is lowered to allow airflow through the nasal tract and the soft palate must move back to its normal position to produce the following vowel [a]. Similarly, when [a] is followed by [m] (in the word "lamb"), the velum will begin to lower itself during the articulation of [a] that it would be ready for the following nasal. The transitions between articulatory gestures are produced by physical system of muscles.

Some aspects of these gestures (e.g. lip contact in bilabial stops) vary less than others (e.g. jaw rotation in vowels.), as a result phonemes have varying influence over a speech utterance. This varying dominance can be used to determine final trajectories of phonemes ranked according to their degree of articulatory constraint (Recasens et al., 1997). Contextual effects of coarticulation have been observed up to seven segments. In summary, coarticulation regards to the extent to which a gesture is realised and also respects the influence of that gesture over a period of a speech act.

Two forms of coarticulation can be separated: Anticipatory (forward) coarticulation, when the vocal tract is preparing for an upcoming important gesture and perseverative (backwards) coarticulation, when the effects of a sound are seen during the production of sound(s) that follow it.

Different researchers identified various coarticulation control models to reproduce the effects of anticipatory and perseverative coarticulation on natural speech and to govern the articulatory movements for a given phonetic target specification. Existing models can be divided into look-ahead models and time-locked models.

Look-ahead models. In look-ahead models an anticipatory coarticulatory gesture begins at the earliest possible time allowed by the articulatory constrains of other segments in the utterance. In Ohman's (1967) model, the time varying shape of the vocal tract is modelled as a vowel gesture where the vocal tract gradually changes shape from V1 to V2, onto which a consonant gesture is superimposed. The consonant has an associated temporal blend function that dictates how its shape should blend with the vowel gesture over time. It also has a spatial coarticulation function that dictates to what degree different parts of the

vocal tract should deviate from the underlying vowel shape, i.e. how context-sensitive the realisation of that segment is for the different articulators.

Ohman's model was originally intended to describe lingual coarticulation in simple VCV sequences by predicting cross-sectional distances in the vocal tract. Reveret et al. (2000) modified Ohman coarticulation control model and defined coarticulation between two vowels in phonetic structures V_1CV_2, V_1CCV_2, V_1CCCV_2 etc.

Time-locked models. In time-locked models the beginning of a gesture occurs a fixed time before the start of the associated segment, regardless the timing of the other segments in the utterance. In Lofqvist's gestural model (Lofqvist, 1990) speech production is modelled as a series of overlapping articulatory gestures. Cohen and Massaro (1993) implemented a coarticulation model based on Lofqvist's gestural theory of speech production. Cohen-Massaro used negative exponential functions to represent the time varying dominance functions and to model articulatory gestures during animated speech. In this approach, a viseme shape is defined as dominance functions that are related with facial measurements. Each articulator (lips, tongue, jaw etc.) has a number of related dominance functions which determines the influence of phoneme over its trajectory. Dominance of phoneme varies within each articulator (i.e. Bilabial plosives exert a greater influence over the motion of the lips than of the tongue).The final mouth shape of analysed phoneme is defined as weighted sum of dominance values for neighbouring phonemes. Fig. 1.8 schematically illustrates the essential idea of the Cohen-Massaro Model.

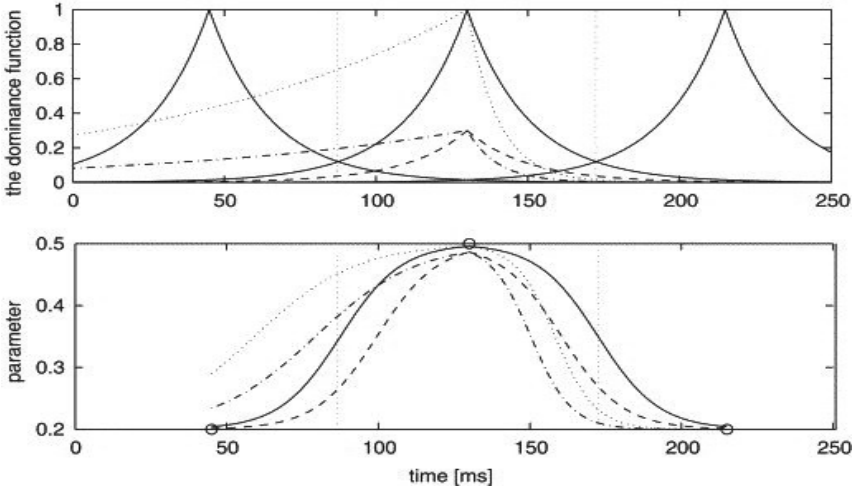


Fig. 1.8. The dominance of three concatenated speech segments over time (top) and the resulting synthesized trajectory of one parameter (bottom)

Basic Cohen-Massaro coarticulation model was improved by various researchers. For example, Cosi et al. (2002) added a temporal resistance function and a shape function for more general cases, such as fast/slow speaking rates; Goff and Benoıt (1996) calculated the model parameter values of the Cohen-Massaro model by analysing parameter trajectories measured from a French speaker; Bevacqua and Pelachaud (2004) proposed an expressive qualifier modelled from recorded speech motion data to make expressive speech animations.

Several limitations of Massaro model have been reported (DeCarlo et al., 1998). For example, certain targets, such as the closure in a bilabial stop, cannot be achieved with this model. Despite this the Cohen and Massaro coarticulation model is the most commonly used by the visual synthesis community.

ANN coarticulation control models. Another way to model coarticulation is the selection of visual speech units from clustered data. (Krnoul et al., 2006) presented method, where several instances of each visual speech unit are stored and clustered using decision trees. Static target vectors, representing points on the face, are used to represent each phone. These target vectors are interpolated to form a continuous track of visual speech. Artificial neural networks (ANN) (Strom, 1997) must be trained to predict articulatory parameter values on a frame-by-frame basis. Input to these networks consists of a series of feature vectors constructed from the time-labelled phoneme sequence using lookup table. Examples of feature vectures are phoneme classes (consonant or vowel), binary place features (bilabial, labiodental, dental, retroflex, alveolar and velar), binary manner features (nasal, fricative, stop, release and voiced), 3-valued vowel features (open, front, rounded and protruded), stress.

Rule-based coarticulation control models. Rule based control schemes have been successfully employed for visual speech synthesis. These coarticulation models use a set of explicit rules to model steady-state properties of pronounced phonemes and parametrically control how these phonemes are fused into connected speech.

Pelachaud et al. (1996) described an implementation of the look-ahead model. Phonemes are clustered into visemes that are classified with different deformability rank, which serves to indicate to what degree that viseme should be influenced by its context. Visemes with low deformability serve as key-shapes that influence the shape of the more deformable ones.

Beskow (1995) proposed another rule–based model, where each phoneme is assigned to a target vector of articulatory control parameters. Some parameter values can be left undefined to allow these targets to be influenced by coarticulation. If a target is left undefined, the value is inferred from context using interpolation. For example, the lip rounding parameter in V_1CCCV_2 utterance (vowel V_1 is unrounded, V_2 – rounded) is unspecified for the consonants C, so conso-

nant targets are determined from the vowel context by linear interpolation from V_1, to V_2.

Beskow (2004) implemented and compared four coarticulation control models according to their perceptual intelligibility (Table 1.1).

Table 1.1. Summary of intelligibility test of visual speech synthesis coarticulation control models (by Bescow, 2004)

Coarticulation Control model	Keywords correct, %
Audio only	62,7
ANN	72,8
Cohen-Massaro	74,8
Ohman	75,3
Rule-based	81,1

The results confirm that all coarticulation control models give significantly increased speech intelligibility over the audio-alone case. It also proves that coarticulation modelling should be integrated in any approach to visualize speech.

1.2.4. Facial modelling for speech animation

Facial modelling techniques for speech animation can be broadly categorised into video-based and model-based, where either 2D video data is used as raw material for producing new frames of animated speech, or some kind of 3D deformable model is used as the basis for image generation. However, the boundaries are blurring, since video-based systems are starting to incorporate elements of 3D techniques and model-based approaches make increasing use of textures and other pixel-based data sources.

Video-based techniques

The fundamental problem for video-based “Talking heads” is how to seamlessly concatenate video sequences. Even the slightest change in head posture or change in facial expression would be very noticeable if occurring from one frame to the next. Thus, pre-processing steps performed by most systems include normalisation of face position and orientation, as well as segmentation of the face into regions that can be treated separately. For example, in the Video-Rewrite system by Bregler et al. (1997), only the mouth-area is processed, and later re-imposed (with new articulation) into the original video sequence.

Cosattoet al., (2000) divided the face into several regions that are stored separately and can be combined with a base head to assemble new animations .

Image sequences for the mouth area are generated using unit-selection techniques known from acoustic synthesis (Hunt et al., 1996) in order to achieve photorealistic visual speech synthesis. The resulting animations are in general of very high quality, but the approach suffers from the same drawbacks as acoustic unit selection systems, i.e. it can produce undesirable results when there are no appropriate units in the database. In addition, when the system is combined with an acoustic unit selection synthesizer, the system becomes sensitive to mismatches between acoustic and visual units.

An alternative to raw concatenation of image sequences can be to build statistical models of the image bitmaps, allowing image generation to be controlled by a compact set of parameters. Such a model was proposed by Brooke (1996). It is based on Principal Component Analysis (PCA), a technique for reducing the dimensionality of a data set by finding a small set of linear combinations of the original dimensions that optimally describes the variance of the original data. When PCA is applied to images, the grayscale level of each pixel represents one dimension, thus an M-by-N grayscale image can be represented as a vector in an MN-dimensional space. For colour images, each pixel represents three dimensions; red, green and blue. Brooke et al. (1998) applied a two-level PCA by dividing the mouth region into 16 sub-blocks of 16x12 pixels, each of which was subject to PCA . Then a second, global PCA was performed on the first 30-50 principal components

Image-based (2D) speech animation require a huge amount of data recorded in a professional studio environment and selection of sophisticated analysis methods to process recorded video of human speech. Since animations are synthesized only from existing data samples and user cannot arbitrarily choose the texture and point of view of the „Talking head", Image-based models do not offer high level of flexibility.

Model-based techniques

In model based synthesis, the facial surface is typically described as a polygonal mesh, usually in 3D. During animation, the surface is deformed by moving the vertices of the mesh, keeping the topology of the network constant. The movement of the vertices is governed by a set of control parameters. The mapping from control parameter values to vertex displacements can be based on a number of techniques, including parameterisation, pseudo-muscular deformation or performance-driven techniques

Parameterization. One of the earliest computer generated „Talking heads" was developed by F. I. Parke's, who developed a rough polygon representation of the head to create a simple animation of the mouth and eyes opening and closing. Four years later, he completed the first parameterised facial model by grouping vertices together to implement specified non-rigid face deformations.

Lots of the present audiovisual speech synthesizers are based on a parametric face model presented by Parke (1982). The model consisted of a mesh of about 800 polygons that approximated the surface of a human face including the eyes, the eyebrows, the lips, and the teeth. 50 parameters were used to control the polygon surface (Beskow, 1996). However, present systems contain a number of modifications of Parke model to improve it and to make it more suitable for synthesized speech. These are usually a set of rules for generating facial control parameter trajectories from phonetic text, and a simple tongue model, which were not included in the original Parke model.

In summary, three-dimensional parametric models can be viewed as a geometric description of the facial surface that can be deformed using a limited set of control parameters and rendered using standard computer graphics techniques.

Muscle-based implementation. In the direct parameterisation approach, there are no restrictions on the type of deformation that the parameters may exert on the model, since they are defined by arbitrary geometric transforms. While this is a powerful approach, it must be used with care since it easily can yield physiologically impossible results. The other popular method to create „Talking heads" is muscle based implementation. The first physically based facial muscle model was developed in the 1980's. Waters (1987) presented „Talking head", where 22 muscle groups were defined and parameterised using Facial Action Coding System (FACS) (Ezzat et al., 2002). Muscles were controlled by a single Facial Action Unit (FAU) parameter, where each FAU parameter deforms the polygonal skin model to mimic the effects of contraction on the skin. This approach allowed a variety of facial expressions to be created by controlling the underlying muscle structure of the face. The muscle based „Talking head" modelling became the basis of commercial animation implementations. For instance, a muscle model, similar to that Waters (1987) proposed, was used to animate a set of 3-D points, which represented skin of computer modelled baby, in a short film „Tin Toy" developed in 1988 by Pixar Animation Studios. A muscle model was used to animate a set of 3-D points, which represented Billy's skin. By "moving" groups of muscles, the animator had control over the aesthetics and choreography of the face. However, this kind of head models requires a highly skilled animator to achieve a believable animation. Besides it is time consuming and expensive animation. Due to difficulties with muscle based implementation, most researchers have found the parametric model more feasible (Beskow, 1996; Le Goff et al., 1996).

Performance-driven implementation. In the performance-driven approaches presented to model facial surface deformation, less attention is paid to the underlying physiology of the face. Rather, just as with the direct parameterisation method, the goal is to model surface deformation directly. But instead of relying

on manual observations, physical measurements form the basis of the parameterisation that is derived using statistical methods, usually some variant of Principal Component Analysis (PCA). Fundamental to these methods are the data capture techniques. Most importantly, data must have high spatial resolution in order to provide reliable estimates for the displacements of all vertices in the target 3D-mesh model; at least several hundred vertices are used in most models, often several thousands.

Kuratate et al. (1998) obtained a set of eight high-resolution 3D head scans with a Cyberware scanner, representing different static speech and non-speech orofacial configurations. A generic 3D facial mesh was fit to the high-resolution data of each scan. Then a PCA was applied to the vertices of the facial mesh, resulting in five linear components accounting for 99% of the variance between the eight key-configurations that can be used to deform the mesh during animation.

Elisei et al. (2001) used manual 3D photogrammetry on video, to track the position of 168 coloured beads glued to a speaker's face. The coordinates of the flesh points were subject to a modified PCA, resulting in six linear orofacial control parameters describing 97% of the variance in 40 recorded key-postures. However, expensive scanners are required for this technique.

Comparison of video-based and model-based animation

"Talking heads" can be compared in automatism, flexibility and realism level. Automatism evaluates processes of model creation and speech synthesis. Flexibility includes characteristics like synthesis speed and possibility to render „Talking head" model from arbitrary views and user-defined textures. Realism evaluates how precisely animated speech matches a natural speech (Cosatto et al., 2003). The strengths and weaknesses of image-based and 3D model-based facial animation techniques are presented in Table 1.2 (Ostermann et al., 2004).

Table 1.2. Comparison of Video-based and 3D models based facial animations in terms of automatism, flexibility and realism

Characteristics	Video-based animation	3D model-based animation
Automatism	low	high
Flexibility	low	high
Realism	high	medium

3D models based „Talking heads" lack realism of visual speech (especially in the mouth area). The mouth zone demonstrates a lot of small wrinkles during natural speech, but it is hard to model those lip movements properly. Moreover,

visualization of the human mouth causes high deformations of 3D models during animation, but the texture-maps of the 2D images remain the same. A lot of current speech animation researches attempt to create methods to improve realism of speech animation in the mouth zone. In the meantime, video-based facial animation techniques achieve greater realism in synthesized videos (Cosatto et al., 2003).

3D „Talking heads" have a much higher automatism level then the video-based „Talking heads", since adaptation of generic 3D head model executed to visualize new human faces is highly automated. Moreover, 3D head models may be rendered from arbitrary views together with user-defined textures. High automatism and flexibility of 3D model-based animations makes it more suitable for translingual speech animation analysed in this thesis.

1.3. Facial animation systems

1.3.1. Face parameterization and scripting languages

Positions of the mouth and tongue must be related to characteristics of the speech signal to generate a naturally looking „Talking head". Multimedia Content description mechanisms are analysed to outline facial movements. Facial Action Coding System (FACS) (Ekman et al., 1978) was one the first attempts to model the low level facial movements by defining a set of coded actions. MPEG-4 (Battista et al., 1999) standard uses a similar idea to describe facial features and movements.

MPEG-4 Facial Animation (MPEG-4 FA) includes Face Definition Parameters (FDPs) and Face Animation Parameters (FAPs), where FDPs define a face by measures of its major parts and features (like eyes, lips and their related distances) as well as FAPs encode the movements of these facial features. MPEG-4 face parameters define a low-level mechanism to control facial animation. Although very powerful, this mechanism lacks higher levels of abstraction, timing control and event management. The MPEG-4 standard also includes extensible MPEG-4 Textual format (XMT) framework (Kim et al., 2000) to represent scene description in a textual format providing interoperability with languages such as SMIL and VRML. It consists of two levels of textual formats. XMT-A is a low-level XML-based translation of MPEG-4 contents. XMT-Q is a high-level abstraction of MPEG-4 features, allowing developers to create the scene description in languages like SMIL and VRML. These descriptions later can be compiled to native MPEG-4 format to be played back by MPEG-4 systems. However, XMT languages do not include face-specific features.

Recent advances in developing and using Embodied Conversational Agents (ECAs) (especially their web-based applications) and growing acceptance of XML as a data representation language have drawn attention to XMLbased markup 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.

After a series of efforts to model temporal events in multimedia streams (Hirzalla et al., 1995), important progress was made in multimedia content description with Synchronized Multimedia Integration Language (SMIL) (Bulterman, 2001), an XML-based language for dynamic (temporal) description of the events in a general multimedia presentation. It defines time containers for sequential, parallel, and exclusive actions related to the presented objects, in order to synchronize the events. However, SMIL does not act as dynamic content description for facial animation or any other specific application.

BEAT (Cassell et la., 2001) is another XML-based system, specifically designed for human animation purposes. It is a toolkit for automatically suggesting expressions and gestures, based on a given text to be spoken. BEAT uses a knowledge base and rule set, and provides synchronization data for facial activities, all in XML format. This enables the system to use standard XML parsing and scripting capabilities. Although BEAT is not a general content description tool, it demonstrates some of the advantages of XML-based approaches.

Multimodal Presentation Markup Language (MPML) (DeCarolis et al., 2002) is another XML-based markup language developed to enable the description of multimodal presentation on the WWW, based on animated characters. It offers functionalities for synchronizing media presentation (reusing parts of SMIL) and new XML elements such as <listen> (basic interactivity), <test> (decision making), <speak> (spoken by a TTS-system), <move> (to a certain point at the screen), and <emotion> (for standard facial expressions). MPML addresses the interactivity and decision-making, but suffer from a lack of explicit compatibility with MPEG-4 (XMT, FAPs, etc).

Virtual Human Markup Language (VHML) (Marriott et al., 2002) is an XML-based language for the representation of different aspects of "virtual humans", i.e. avatars, such as speech production, facial and body animation, and emotional representation. It comprises a number of special purpose languages, such as EML (Emotion Markup Language), FAML (Facial Animation Markup Language), and BAML (Body Animation Markup Language). In VHML, timing of animation elements, in relation to each other and in relation to the realisation of text, is achieved via the attributes "duration" and "wait".

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 to design Face Modelling Language (FML) which allows re-use of

existing XML tools and products (Arya et al., 2004). 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 animation scripts can be 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, meaningful actions and even stories.

Table 1.3 summarizes scripting languages that may be used for facial animation and their supported features.

Table 1.3. Content description languages and their features

Feature	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

The only MPEG-4 compatible face parameterization language is FML, which also has a high-level face components and behavioural modelling.

1.3.2. Current non-commercial facial animation systems

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

Wang et al. (2007) 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 et al. (2005) 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. (2007) 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 (2007) proposed a facial animation framework compatible with the MPEG-4 standard called „iFACE". „iFACE" allows interactive non-verbal scenarios through FML scripting language (Arya et al., 2007). FML allows both parallel and sequential description of face actions that include talking, expressions, head movements and low-level MPEG-4 parameters, so „iFACE" is very suitable for translingual viseme-driven speech animation.

1.4. Facial animation evaluation methods

Synthesized „Talking heads" can be used in a wide range of applications. However effective scientific methods are required to evaluate the deficiencies of these models.

The quality of synthetic facial animation, produced solely from speech can be measured using various approaches:

- Visual comparison of animation parameters in synthesized and recorded speech (Theobald et al. 2003) (Cohen et al. 2002).
- Forced choice approach where the viewer is asked to state whether a displayed animation is real or synthetic. The tests may be thought as a kind of *Turing Test* for facial animation ((Ezzat et al., 2002) (Hack and Taylor 2003), (Geiger et al., 2003)).
- Measurement of the ability of a synthetic „Talking head" to improve the intelligibility of speech in a noisy environment (0uni et al., 2003).
- Subjective evaluation (Bregler et al., 1997; Cosatto and Graf, 2000; Ezzat and Poggio, 1998), where collected comments of test participants are used to estimate example videos of animated speech.

1.4.1. Advantages and disadvantages of evaluation methods

Visual comparison in synthesized and natural speech involves comparing of the trajectories of the parameters in synthesized speech animation with trajectories of parameters defined in speech obtained from a real speaker. This method can be defined as the objective evaluation. It provides insight into an algorithm's

strengths and weaknesses together with a quantitative measure of a system's overall effectiveness (Cosker et al., 2004). However, it doesn't offer any results for the perceptual quality of an animation (the realism of the video). Human speech exhibits natural variation, ever since we never pronounce the same things in exactly the same way, so a synthesizer shouldn't be expected to generate a parameter sequence *exactly* as it appears in the natural language. In this method differences between ground-truth and synthesized speech are perceived as errors, although they can also be attributed as variation observed in natural speech. This variation cannot be quantified using purely objective measures, so synthesised „Talking heads" require subjective evaluation (Theobald et al., 2008).

An obvious method to evaluate synthesized visual speech subjectively is to adopt a Turing-type test, where the viewers are asked to state whether a displayed animation is real or synthetic (Geiger et al. 2003, Ezzat et al. 2002). If the animations are indistinguishable from real video, then the chance of correctly identified synthetic animation is 50/50. A drawback of this forced choice approach is that the participants can have a *prior opinion* about the analysed animation and it may influence their decisions. Any artefacts picked up in the animations (e.g. texture flicker or incorrect coarticulation) can impact viewer opinion. Another drawback of this method is that good testing results can be obtained by randomly selecting either "real" or "synthetic" choices for each clip when participant is bored or uninterested to behave. Moreover, *video realistic* sequences that are indistinguishable from real video should be generated for this test. Some creators of speech animation systems (e.g. (Massaro, 1998)) seek to develop the realistic model of speech, but they do not strive to produce the realistic sequence of images. Therefore, the forced choice approach cannot be applied to the full spectrum of speech animation systems.

Measurement of the ability of a synthetic „Talking head" to improve the intelligibility of speech in a noisy environment gives a good indication of the quality of generated animation. Cosker et al. (2004) measured the performance of their visual synthesizer using experiments based on the McGurk Effect (McGurk et al., 1976). McGurk and MacDonald noted that auditory syllables such as /ba/ mixed with visual stream of syllables like /ga/ were perceived as an entirely different syllable, e.g. /da/. So, they argued that when information from both visual and auditory sources is available, human speech perception system finds the best-fit solution for the conflicting audio and visual information (*McGurk Effect*).

The idea of this subjective evaluation method is to present monosyllabic words with concurrent and inconcurrent audiovisual stimuli in both synthesized and real video conditions. The viewers are asked to observe McGurk Effect for the real and synthesized stimuli, one video clip at a time. For each video participants are asked to name the word they hear while watching. If the synthesizer

generates inaccurate speech gestures, the expected response (predicted from McGurk) will not match the given response. Expected responses are provided and checked in a real video. The disadvantage of this method is that only isolated monosyllabic words are used in this method, so it tells us nothing about longer-term dynamics that relate to the overall speech naturalness. Moreover, synthesizers may look realistic and entirely plausible over single words, but longer term coarticulation effects may be less well captured and analysed with this evaluation method.

Most experiments of subjective evaluation attempt to collect comments on different aspects of visual speech synthesiser and to sum-up these results for the final estimation. All components of the system can be considered simultaneously or evaluated separately. For example, the point-light method (Bailly et al., 2002) separates the speech model from the appearance of the face to identify inadequacies of the system. On the other hand, a dense coverage of point-lights scattered about the face is required to accurately capture subtle movements of the lips, jaw and cheeks. More often, animation systems are evaluated both in terms of intelligibility (Massaro et al., 1998; Bailly, 2002; Benoît et al., 1998; Fagel et al., 2007) and naturalness (Theobald et al., 2004; Fagel, 2007-2). These examples of subjective evaluation provide subjective information on the overall quality of facial animations, but leave no means of comparison with other systems.

1.4.2. Challenges in evaluation of different speech animation systems

Comparison of the overall quality of one speech animation system against other ones is complicated, because of a number of issues. 6 main problems can be recognized:

1) There is no common evaluation scheme of the animated speech. The test data to be synthesized and the testing methodology differ between various studies.
2) Different evaluation metrics are used to judge the performance. Some systems are evaluated only objectively (e.g. RMS error in geometric features or articulation parameters), some are also evaluated subjectively (e.g., naturalness, intelligibility, etc.).
3) The structure of the data used to train different speech animation systems usually varies. For instance, the number of sentences and their phonetic makeup balanced for diaphones or triphones coverage may differ in the adopted language of visual speech synthesizer.
4) Different speakers are employed to capture data for the experiments and some of them are better than others in terms of clarity of articulation, speaking style and etc.

5) Viewing conditions of the evaluation experiments are not the same. Some tests adopt strict viewing conditions with controlled lighting and sound proofed listening conditions, whilst others are less formal.
6) Distinct speech animation systems use differently looking face (head) models. Some strive for videorealism, whilst others adopt human-like graphics-based approaches. The uncanny valley effect (MacDorman, 2006) states that when the synthetic face model is made more human alike, observer's emotional response is becoming increasingly positive and empathic, until the point, where human replicas behave in a way very close to, but not perfectly the same as, human behaviour. Beyond this point, the viewer feels a strong revulsion. However, as the human face replicas continue to become less distinguishable from a real human being, the emotional response becomes positive once more. Fig. 1.9. shows „the 'Uncanny Valley' affect and the positivity of responses to human replicas.

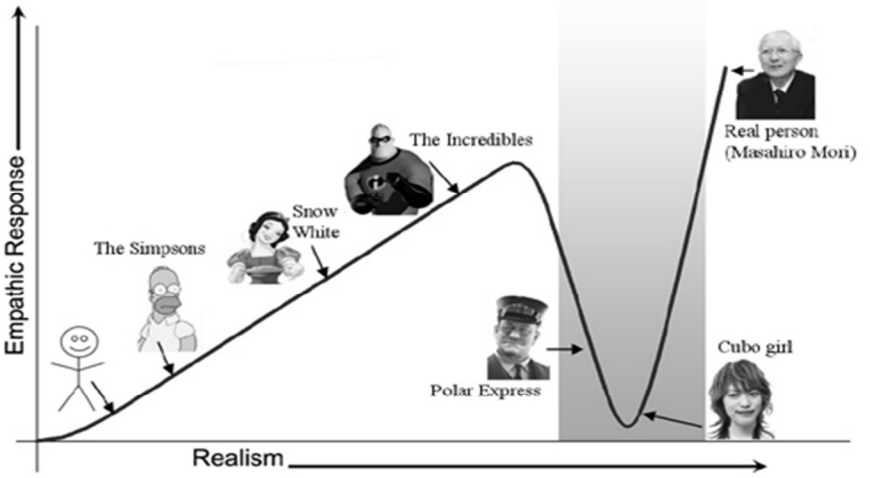


Fig. 1.9. "The uncanny valley" effect. Hypothesized emotional response of human subjects is plotted against anthropomorphism of a synthetic face. The uncanny valley is the region of negative emotional response towards syntetic faces that seem "almost human"

Recently, researchers in auditory speech synthesis have sought to overcome these issues by running evaluation in the form of a competition (i.e. the Blizzard Challenge (Black et al., 2007) and LIPS2008: Visual Speech Synthesis Challenge (Theobald et al., 2008)). These competitions have proven to be popular and provided framework for unifying and standardising evaluation between dif-

ferent research groups. Since English language is used to train visual speech synthesizers that participate in these experiments, data of the competitions (the acoustic files, the list of sentences used and the phonetic transcriptions in various formats) cannot be reused for Lithuanian speech animation evaluation. However, approach similar to LIPS2008 was used for subjective evaluation of Lithuanian speech animation framework.

1.5. State of art in Lithuania

Speech animation is based on processing and synchronization of 2 different signals: acoustic and visual. Both of these signals are highly dependent on the phonetics of specific language.

In 2006 Kaukėnas et al. presented audiovisual speech recognition system, which consists of audio and video detection in the recorded video of talking man. This system uses viseme extraction technique to recognized text what was pronounced in a video file. Features from video and acoustic streams later are integrated to make a final decision about the recorded speech. However, viseme classification into viseme classes was not considered in this research. Classification of Phonemes using Sound and Visual Information was analysed by Balbonas et al., (2005).

8 Lithuanian viseme groups were manually detected based on subjective observation of the places of articulation during the pronunciation of phonemes. Viseme classes relate to articulators that are visibly contrastive. For instance, two bilabial plosives /p, b/, dental consonants /d, t, s, z, n, dz, l/ or velar /g, k, h, ch, j/ are defined as three separate visemes. Moreover, differently looking vowels and diphthongs /a, i, e, ė, ai, ei, ie/ was classified as single viseme. Based on this classification 5 visemes for two Lithuanian words animation were created and used in Master thesis of Zailskas (2011). Vectored 2D human face was created to make speech animations without sound input. More detailed examination of visemes classification is needed to cerate naturally looking speech synchronized 2D or 3D „Talking head".

Speech visualization techniques for Lithuanian language are just beginning to be explored, however Lithuanian speech recognition are much more widely analysed.

Current Lithuanian speech recognition engines. Accuracy of speech recognition engine together with correctness of phonemes to visemes mapping notably influences the output of animated speech and the realism of animated speech.

In 1999, Rudzionis et al. described the phoneme recognition system based on regularized discriminant analysis. This system discriminated a subset of Lithuanian phonemes in a fixed context.

Later investigations on Lithuanian speech recognition are based on different techniques: dynamic time warping (Lipeika et al., 2002; Tamulevicius et al., 2004), Hidden Markov modeling (Laurinciukaite, 2004; Raškinis et al., 2003); pure Artificial neural networks (ANN) (Filipovic, 2003) or mixed HMM/ANN technology (Filipovic et al., 2004).

Lipeika et al. (2002) described an isolated-word speech recognition system based on a dynamic time warping (DTW) technique. The system extracted linear predictive coding (LPC) features from a speech signal and used speaker adaptation techniques. It was trained to recognize Lithuanian spoken digits and had the vocabulary of 12 words.

Raskinis et al. (2003) described the process of building Lithuanian HMM speech recognizer for the medium vocabulary (>100 words) isolated-word speaker-independent speech recognition. They used Hidden Markov Model (HTK) toolkit (Young et al., 2000) to develop the triphone single-Gaussian HMM speech recognition system based on Mel Frequency Cepstral Coefficients (MFCC). Hidden Markov model's parameters were estimated from phone-level hand-annotated Lithuanian speech corpus. This speech recognizer was evaluated on VMU Speech Corpus and performed at 20% word error rate on 3 out of 4 speaker independent recognition tests.

Hybrid HMM/ANN speech recognition system also based on Mel Frequency Cepstral Coefficients (MFCC) verified noticeably better results on word recognition performance (Filipovic et al., 2004). CSLU Toolkit was used to develop the hybrid HMM/ANN speech recognition system, which combines artificial neural networks (ANNs) and hidden Markov models (HMMs). In this ASR system a fully connected three-layer neural network (a multi-layer perceptron) was trained by conventional stochastic back-propagation algorithm to estimate the probability of 115 context-independent phonetic categories.

The newest HMM-based Lithuanian speech recognition engine (Šilingas et al., 2006) was created with HTK toolkit and the original Java based software, which was developed for syllabication, frequency analysis, deriving dynamic phone sets, phone-level recognition accuracy analysis, and transformation of pronunciation vocabulary, phonetic transcriptions and models. This speech recognition system was evaluated on long duration speech corpora LRN0, LRN0.1 and LRN1.

Speech corporas. The most modern speech recognition methods are based on statistical models that require large amounts of training and test data for evaluating and tuning of parameters (Laurinciukaite et al., 2006). Specialized speech

databases, called speech corpora, are collected to analyse specific parameters of analysed language.

Different corpora are collected because of the high variability of speech signals in vocal tracts of speakers, environments, communication channels and style of speech. *Specialized* corpora are oriented to particular features of speech: dialects; male/female/children; spontaneous speech; reading of the text; individual words; commands and continuous speech, while the *universal* corpora cover many common features of speech. Larger corpora provide a better representation of language variability, but their use depends on how well they are constructed.

The most valuable characteristic of a speech corpus for researchers is a comprehensive annotation of speech signals, i.e., time-aligned annotation of a speech signal on different levels: phoneme, word, and syllable. Typically annotations vary from phonetic labelling of segments (phonetic transcriptions as in TIMIT (Lamel et al., 1998)) to labelling of some semantic category (Kelly et al., 2006).

Different Lithuanian universities and scientific laboratories constructed and used their own speech corpora. Major Lithuanian speech corpora are LTDIGITS (Rudžionis et al., 2002), Vytautas Magnus University (VMU) Speech Corpus (Kulešienė et al., 2005) and LRN 1 (Laurinciukaite et al., 2009). All of them have phoneme and word level annotations.

LTDIGITS has been created as the Lithuanian speech database and a tool that could be used for analysis of discrimination of nasals in various vowel environments. 225 female and 125 male speakers were observed in the laboratory (acoustic camera) conditions, to collect data for the LTDIGITS corpora. It contains Lithuanian computer control words and Lithuanian continuous digit sequences. Important feature of LTDIGITS is collected set of CV (consonant-vowel) clusters containing nasals together with various vowels (MA, NA etc.) and same syllables in continuous phrase. Both the phrases to words and word to phone marking and labelling procedures were developed to process LTDIGITS. The labelling results were saved in TIMIT style and in the modified formats where only word/phone boundaries were kept.

VMU Speech Corpus is universal annotated Speech Corpora of the Lithuanian language, which was built upon 4 speakers (2 males and 2 females) voice records, where each record has the same set of 740 isolated words. The aim of the corpus was to give possibility to use popular tools like HTK, MBROLA etc. for the investigation of Lithuanian spoken language. The vocabulary of the corpus was chosen to include all distinct and independent Lithuanian sounds such as phonemes and phoneme clusters (275 phonetic units in total). Information about beginning and the end times of the word, phonetic unit boundaries of the word and phonemes was used to annotate each sound record with words, phonet-

ic units and sound level. This information is saved into 275 annotated "TextGrid" files. Freeware software "Praat" was used to annotate the corpus.

Both LTDIGITS and VMU Speech Corpus support detailed investigation of speech acoustics. However, these speech corpora are of medium duration (1-6 hours) and do not contain enough continuous speech data from multiple speakers that are necessary for constructing speaker-independent large vocabulary continuous speech recognition systems. Speech corpus LRN 0 and its later versions LRN 0.1 and LRN 1 (Laurinciukaite et al., 2009) were developed to solve this problem. It was created at the Institute of Mathematics and Informatics (IMI) in cooperation with VMU. This corpus contains over 21 hours of speech samples from the news broadcasts by Lithuanian Radio in 2003-2004. Speech corpus was divided into training (18 hours, 13 341 sentences), development (37 minutes, 736 sentences) and evaluation (2 hours, 1775 sentences) data sets. LRN 1 is accompanied by time-aligned word level transcriptions lexicon and contains 28 386 entries. HTK tools, model (speech data) realignment were major methods for automatic annotation of speech on the word level.

Two main steps were used for the realignment of this speech corpus:

1. Training of hidden Markov models (HMMs) based on a selected phoneme set.
2. Realignment of data in the recognition process.

Since HTK tools have limited options for a comprehensive investigation of speech data in manual setting of boundaries, Freeware software "Praat" was additionally used to ensure that annotations of speech data had correct boundaries of the word. Validation of the speech corpus was performed on the new Lithuanian speech recognition system (Šilingas et al., 2006), which is based on Hidden Markov model (HMM) methods. After the revision, we decided to include this speech recognition system in our proposed framework for Lithuanian speech animation.

1.6. Conclusions on Chapter 1

In this chapter audiovisual speech generation methods and facial modelling, suitable for speech animation were reviewed. Coarticulation modelling and speech processing for speech animation were revised also.

One of the main tasks of this thesis is to propose speech visualization framework that would be suitable for translingual speech animation and guarantee naturalness and intelligibility of visualized speech. „Talking head" modelling technique meeting these requirements was revealed (Fig. 1.10.) based on the

survey of audiovisual speech generation, speech processing and facial modelling methods.

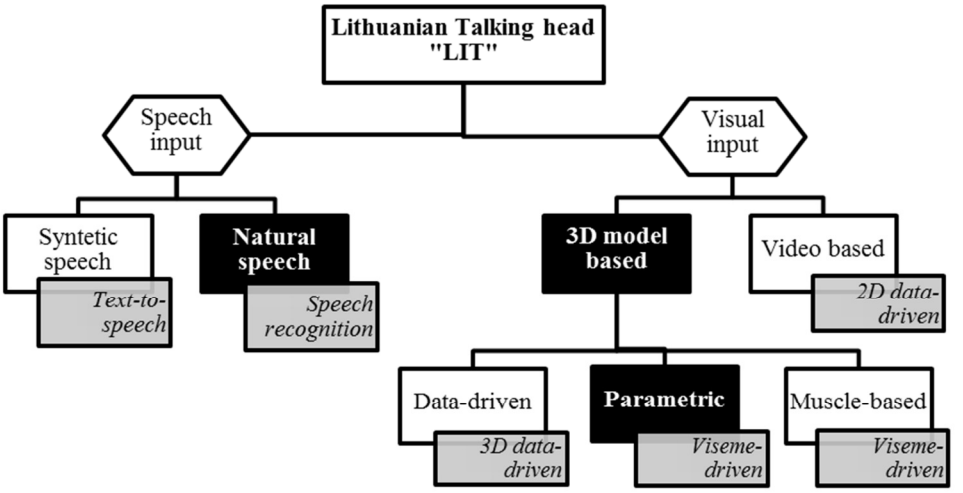


Fig. 1.10. Selection scheme for „Talking head" modelling technique that is the most suitable for Lithuanian „Talking head", which is based on translingual adaptation. Black rectangular mean the best choice in that level

Summing up, Lithuanian „Talking head" that we called „LIT", should be driven by natural speech and should be animated using 3D face model, which is created using parameterization technique. Viseme-driven speech animation should be employed since intuitive management of visual speech synthesis and possibility to re-use viseme models for a new „Talking heads" is very important for translingual speech animation. However, this approach requires challenging and careful manual work to identify phoneme-to-viseme mappings.

As of date, neither a single classification method for visemes, nor a standardized viseme set exist for different languages. Nonexistence of standardized viseme set is the reason why identical phonemes are associated with different viseme classes in various languages.

Efficiency of audio to visual mapping at phoneme level exceptionally depends on the speech processing module. Common automatic speech recognition methods that use HMM models can be applied to process speech in different languages; however each language has its own unique features. It results that individual speech recognition engines must be created for distinct languages. Analysis of speech recognition systems for Lithuanian language was completed

to select the most appropriate Lithuanian speech recognizer for audio visual mapping at phoneme-level.

Coarticulation modelling considerably improves intelligibility of the „Talking head" and should be integrated in any approach to visualize speech. Implementation of rule-based model demonstrated the best intelligibility result, despite the fact that it is limited by people's incomplete understanding of coarticulation effect and their inability to build a full set of rules for phonemes, that are influenced by neighbours. Data driven Ohman coarticulation control model, which is defined to model coarticulation between two vowels tied for the second best result. However, Lithuanian language has specific phonetic features that must be included in the applied coarticulation model in order to achieve the higher naturalness of animated speech.

Survey on current non-commercial facial animation systems with face parameterization and scripting languages was carried out to identify existing speech visualizer that is based on phonetics in other language, but can be adapted to animate Lithuanian language. Moreover, „Talking head" model implemented in this engine should meet the requirements identified for Lithuanian „Talking head" „LIT". Since there is no relevant speech animation systems proposed for Lithuanian language, English viseme-driven speech animation system are going to be used to animate recorded Lithuanian speech.

Estimation of animated speech significantly depends on the human observer, so methods that can be used to evaluate Lithuanian „Talking head" was also reviewed in this chapter.

2

Framework for Lithuanian speech animation

The primary goal of this thesis is to propose speech animation architecture suitable for Lithuanian Speech Animation. Invention of new speech animation system is time overwhelming and requires considerable financial support and specific knowledge, so it is very important to explore the possibility to use the speech animation system of the base language to animate the new language in which the video has to be synthesized.

State of art presented In Chapter 1 was carefully reviewed to define properties of „Talking head" that could be adapted to animate Lithuanian speech. The revealed requirements are:

- Parametric 3D head model, which is easily adjustable and has an option to apply new textures.
- Viseme-driven speech animation generation with possibility to edit its phoneme to viseme mapping table. Audio to visual mapping should be resolved on phoneme level.
- Speech coarticulation effects on particular phonemes and visemes can be described in speech synchronization algorithm, using struc-

tured content description language compatible with MPEG-4 parameters.
- Speech processing, animation control and head adaptation modules are independent from each other and can be called individually.
- Possibility to edit defined visemes and to create the new ones in a relation with MPEG-4 standard.
- Possibility to integrate a new head model.

„Talking heads" can be driven by input text or input speech. Speech-driven models utilize acoustics (and phonetic alignment) of natural human speech and makes language animation more realistic, but more complicated to produce, since speech recognition is very challenging task. Text-driven „Talking heads" employ synthesized voices and synthesized head models to visualize speech. Synthetic voice decreases the naturalness of „Talking head", therefore speech-driven facial animation is created in this thesis.
The ideas presented in this Chapter were published in (Mažonavičiūtė, Baušys 2010; Mažonavičiūtė, Baušys 2011).

2.1. „iFACE" – speech animation system

Taking into account the requirements listed above, we have chosen „iFACE" (DiPaola et al., 2007) as the the base element for the proposed Lithuanian speech animation framework. It was created to animate English language. „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" includes an off-line design environment, *iFaceStudio*, for creating animations and configuring the head objects, and a wrapper control, *FacePlayer*, that can be easily used in web forms and similar GUI applications. Hierarchical 3D head model can be used to control facial actions from vertex to feature-group levels Scripts based on structured content description language FML (Arya et al., 2007).can produce speech animation by accessing the object methods and properties.

„iFACE" is the free/open source speech animation system written in C# programming language. However, English speech recognition engine, that can be used to generate phonetic alignments from the recorded speech file automatically, is not open for editing. The overall „iFACE" architecture is presented in Fig. 2.1.

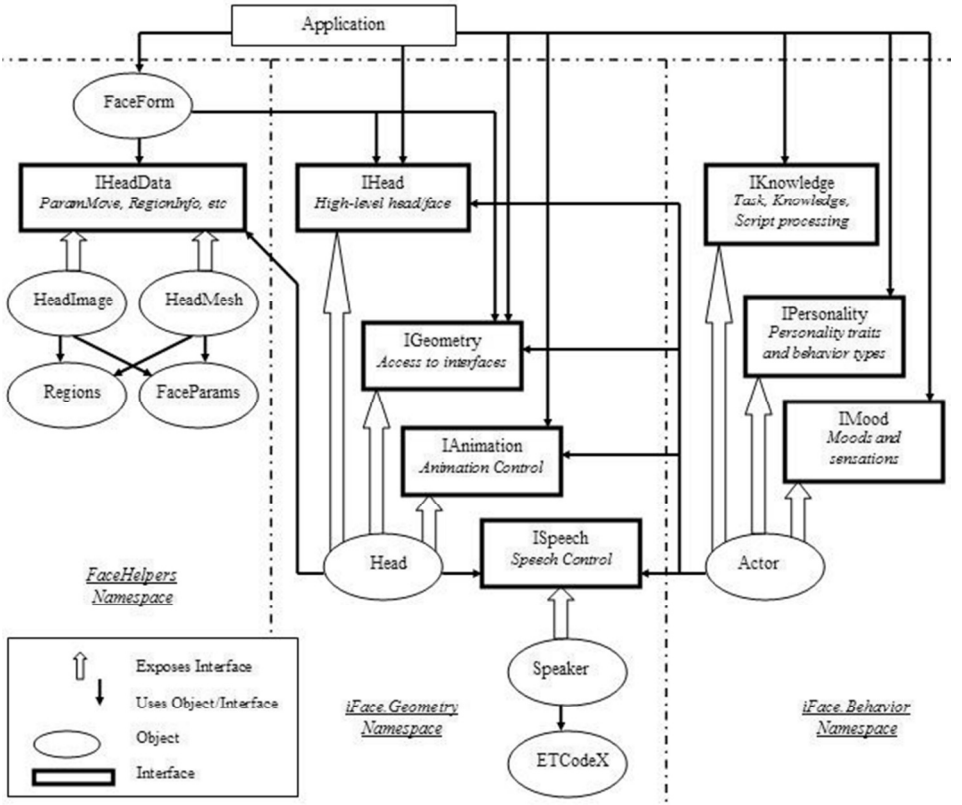


Fig. 2.1. Architecture of „iFACE" speech animation system, originaly created for English language

2.1.1. Geometry of the head model

The actual head data is stored in an object called HeadData (***IHeadData*** interface). It includes *Regions* and *FaceParams* objects. Regions are small groups of points that usually move together. *FaceParams* has members for FDP and FAP, where serialized FDPs define the topology of the head (feature points). FAPs are used in runtime for animation and basically they not serialized. However FAPs in Group-1 (viseme and expression) are serialized combinations of other FAPs.

Geometry of the head model is a hierarchy of abstraction layers on top of HeadData. These layers are **Main** (i.e. the whole head), **Components** (The base head, Eyes, Ears, Brows, Nose, Mouth and etc.), **Features** (i.e. detail lines around main parts of Components), **FeaturePoints** (MPEG-4 FDPs, FACS et.)

and **PhysicalPoints** (actual Mesh Vertices or Image Pixels). Relation between these layers is presented in Fig. 2.2.

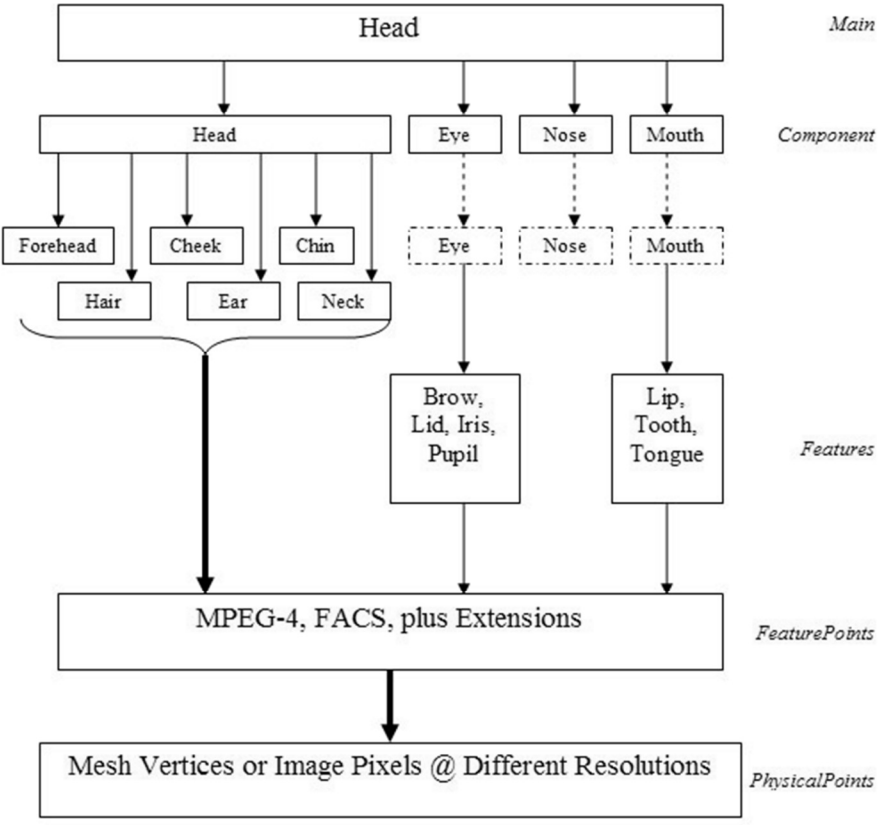


Fig. 2.2. Hierarchy of abstraction layers that defines geometry of the head model used in „iFACE" speech animation system

Head actions (moving and resizing) are defined for Components and Features which in turn affect the related FeaturePoints and their Regions. All action requests are primarily sent to Head object (***IHead interface)***. If necessary Application can ask Head to create and use lower level objects (e.g. to use a Component for actions with wide area of influence or send the action request directly to FeaturePoints). Head object is the main object in Geometry module, which controls the HeadData objects by accessing IHeadData interface.

Head object also implements and exposes ***IAnimation*** for animation control. It creates default speech controller object that implements and exposes the

ISpeech interface. Speech module is part of Geometry namespace but is implemented as a separate DLL to allow defining other speech controllers. Default speech controller uses *ETCodeX ActiveX*, a non-managed code.

2.1.2. Behavior modelling and Face Modelling Language

The Behavior module includes three components, that are designed as components around the Geometry and has their own interfaces for access by the application programs. These components are:

1. **Knowledge.** Behavioural rules, stimulus-response association, and required actions are encapsulated into Knowledge. It receives the input script with external events and holds the rules of interaction. All of these are applied to Geometry in the form of parameters at the appropriate layer of abstraction.
2. **Personality.** Different characters can learn and have the same knowledge, but their actions, and the way they are performed, can be different depending on individual interests, priorities, and characteristics. Personality encapsulates all the long-term modes of behaviour and characteristics of an individual. It suggests facial gestures and states based on the explicit actions requested by Knowledge. Personality can be configured through input scripts or interactively.
3. **Mood**. Certain individual characteristics are transient results of external events and physical situation and needs. These emotions (e.g. happiness and sadness) and sensations (e.g. fatigue) may not last for a long time, but will have considerable effect on the behaviour.

Hierarchical geometry combined with Knowledge, Personality and Mood components is used to define facial animation in „iFACE" (Fig. 2.3)

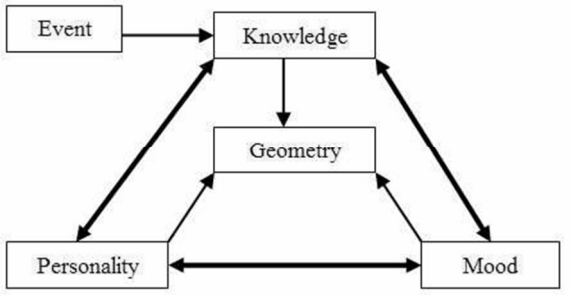


Fig. 2.3. Parameters spaces that determine communicative behaviour of the face in „iFACE" system

For example, if the script requires a piece of speech, Knowledge translates this to a set of phonemes and visemes and their timing, so the Geometry can animate the face. Meanwhile, Personality suggests certain head movements, facial gestures, visemes and expressions that are attributes of the chosen character's personality; Mood applies a base facial state to all the facial actions.

Face Modeling Language. IKnowledge interface is responsible for processing the input scripts written in *Face Modeling Language (FML)*.

Face Modeling Language (FML) is a Structured Content Description mechanism based on eXtensible Markup Language (XML). The main ideas behind FML are:

a) hierarchical representation of face animation (from frames to simple moves, to meaningful actions, and finally stories),

b) timeline definition of the relation between facial actions and external events (parallel and sequential actions, and also choice of one action from a set based on an external event),

c) defining capabilities, behavioural templates, and models (FML is independent of the type of model but provides means of defining it),

d) compatibility with MPEG-4 (MPEG-4 FAPs are supported explicitly and FDPs implicitly by general-purpose model definition mechanisms),

e) compatibility with XML and related web technologies.

FML time containers allow different temporal relations between facial actions, that can be combined both in parallel and sequential ways. Also there is possibility to select one option among a group of facial actions.

2.2. Automatic phonemes recognition for the proposed framework

2.2.1. Specifics of Lithuanian phonetics

Speech animation strongly depends on the chosen speech recognition system. Analysis of language phonetics is significant for speech recognition. Phonetic features – phone duration, softness, diphthongs, assimilations, and specific lexical stress system – need to be addressed in acoustic model of speech recognition system.

Lithuanian is considered to be the most archaic Indo-European language. It possesses features that are difficult to address applying standard speech recogni-

tion methods. The review of Lithuanian language properties that are relevant to speech recognition is presented in (Šilingas et al., 2003)

Lithuanian also has a specific lexical stress system. A stressed syllable may have a rising pitch called circumflex, or a falling pitch called acute. The rising pitch stress can also be short or long. For inflective parts-of-speech the stressed syllable and lexical stress type may change according to inflection. The lexical stress may be critical to discriminate two words, e.g. "šauk" (Imperative of verb "shoot") and "šauk" (Imperative of verb "shout") are discriminated only by the difference between rising and falling pitch stress. Therefore lexical stress information needs to be incorporated both into pronunciation vocabulary and the set of phones used as the basis for acoustic modelling units.

Lithuanian phonemes have some additional features as well. Some pairs of vowels are considered as a single unit, *diphthong,* and some vowel-consonant pairs are also considered as a single unit, *mixed diphthong.* The lexical stress in diphthongs and mixed diphthongs may fall either on the first or second component. There is a question whether the diphthongs and mixed diphthongs should be modelled as single models or should be decomposed in speech recognition systems since some of them, especially stressed mixed diphthongs, may be rather rare.

Lithuanian vowels may be short or long in *duration,* which also needs to be reflected in models. Lithuanian consonants can be *soft* or *hard* depending on the following vowels and consonants. The *softness* of consonant is determined by its nearest context phones in most cases, thus it remains a question whether it should be directly modelled in the BPS or indirectly modelled by the phone context information in contextual models.

There are also multiple effects of Lithuanian phone co-articulation. The two of the most importance for speech recognition are the assimilation of consonants and softening of consonants preceding specific vowels.

2.2.2. Framework for Lithuanian speech recognition module

Continuous speech recognition system, which is used for Lithuanian speech recognition was presented by Šilingas et al. (2006). Modified statistical Bayesian framework (Young, 1995) for HMM-based continuous speech recognition was implemented to form a mixed acoustic modelling units set. Framework, which is based on phonemes and syllables, incorporates significant phonetic features of Lithuanian language.

The structure of the modified system is presented in Fig. 2.4. (Šilingas et al., 2006). Newly created computation modules are denoted as black boxes, standard computation modules as white rectangles, computed artefacts (models, vocabu-

lary and transcriptions) as grey rounded edge rectangles, and the speech corpus as a black data store symbol.

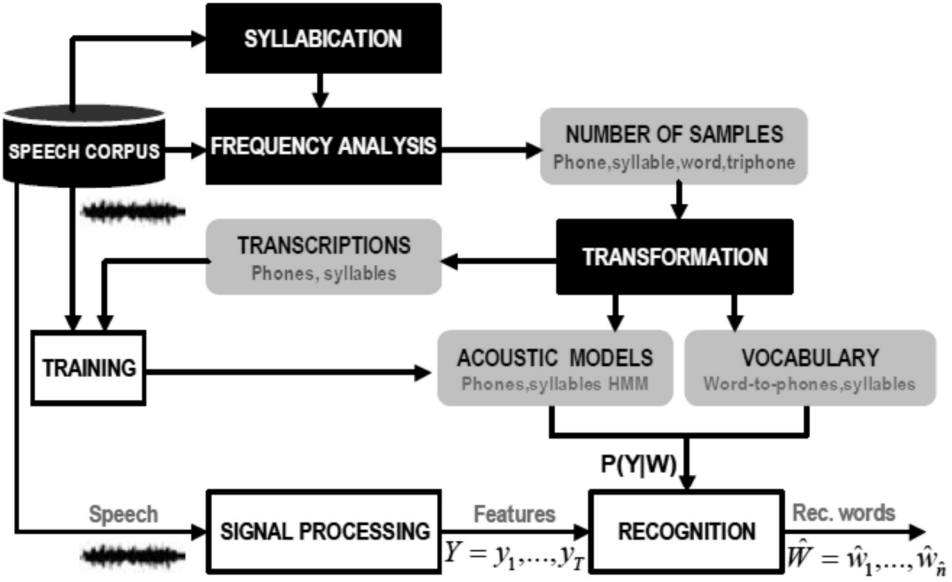


Fig. 2.4. Continuous speech recognition framework for Lithuanian speech recognition

The following modules were introduced in this framework in order to support methodology for forming mixed modelling unit set and choosing complexity for each model:

1. *Frequency analysis* module, which counts training samples for various phonetic units;
2. *Syllabication module*, which forms syllabic pronunciation from available phone pronunciation dictionary;
3. *Transformation* module, which is dedicated to transform pronunciation vocabulary, training data transcriptions and acoustic models. Two major algorithms are applied in this *module*. First algorithm is used to include the most frequent syllables into base phoneme set of vocabulary and acoustic models. The second one is introduced to get phonemes and syllables transcription including linguistic annotation: stress, consonant softness, mixed diphthongs and affricates.

The proposed theoretical speech recognition system was implemented using HTK tools (publicly available HMM modelling toolkit) (Young et al., 2003). HTK toolkit is used for extracting features from speech waveforms, training HMMs and evaluating recognition performance.

Additional software was developed for syllabication, frequency and phoneme-level recognition accuracy analysis, models description, phonetic transcriptions, transformation of pronunciation vocabulary and development of dynamic phonemes sets. It was programmed with Java programming language.

Speech corpus. Speech corpus LRN0 (Lithuanian Radio News version 0) containing about 10 hours of speech (Table 2.1.) was used for the experiments to find the most efficient static phoneme set for Lithuanian speech recognition.

Table 2.1. Major LRN0 characteristics

Criterion	**LRN0 Characteristics**
Speech type	Continuous
Speech content	Read broadcast news
Annotation	Orthogonal word-level transcriptions
Number of speakers	23
Sampling	11 kHz
Quantization	16 b
Channel	Mono
Training data	9+ h (6564 sentences)
Development test data	2 min. (50 sentences)
Evaluation test data	14 min. (360 sentences)
Full vocabulary	18000+ words
Evaluation vocabulary	5500 words
Phoneme set	74 simple phonemes, 156 diphthongs, 3 pseudo phonemes (including softness and lexical stress annotation)

This phoneme set was later reused to create the new versions of this speech corpus (LRN 0.1 and LRN 1). The effectiveness of this phoneme set was proved in additional experiments (Šilingas et al., 2004). For all the experiments, training of the HMMs and recognition was performed using conventional Baum-Welch and Viterbi algorithms respectively.

Investigations of acoustical modelling for Lithuanian continuous speech recognition were based on simple and contextual phoneme set variation. From the results of the executed experiments the following main conclusions were defined (Šilingas et al., 2004):

a) Modelling softness and accent information provides better recognition performance using simple phonemes;

b) Similar three state left-to-right HMM topologies should be reused for both simple phonemes and diphthongs;

c) Since mixed diphthongs are relatively rate, splitting them into two separate phones helps to minimize phone/ triphones set size and does not result in significant decrease in recognition rate using phones and provides significantly better results using triphones;

d) Softness information is implicitly modelled by context information in triphones, thus triphones can be constructed from phone sets without softness information;

e) Lexical stress annotation should be included in the basic phoneme set for Lithuanian language.

f) Analysis of phonemes error rate (PER) for every phoneme (Table 2.2.) showed that diphthong and vowel recognition is more problematic, so models for every diphthong and vowel have to be studied individually.

Table 2.2. Phoneme error rates (PER) for the simplest Lithuanian phoneme set

Nr	*Phone*	*PER, %*	*Nr*	*Phone*	*PER, %*	*Nr*	*Phone*	*PER, %*	*Nr*	*Phone*	*PER, %*
1.	ui	25.00	**11.**	ie	5.99	**21.**	b	3.88	**31.**	m	1.15
2.	uo	13.91	**12.**	i	5.79	**22.**	p	3.67	**32.**	g	1.14
3.	i:	12.53	**13.**	e	5.68	**23.**	au	3.64	**33.**	o	0.95
4.	a:	10.88	**14.**	u	5.61	**24.**	v	3.59	**34.**	zh	0.76
5.	u:	10.03	**15.**	o:	5.23	**25.**	r	3.05	**35.**	ch	0.0
6.	eh:	8.88	**16.**	a	4.92	**26.**	t	2.59	**36.**	eu	0.0
7.	e:	8.86	**17.**	f	4.35	**27.**	k	1.98	**37.**	h	0.0
8.	ai	7.11	**18.**	d	4.15	**28.**	n	1.90	**38.**	z	0.0
9.	ei	7.08	**19.**	l	3.91	**29.**	w	1.82			
10.	j	6.04	**20.**	s	3.90	**30.**	sh	1.25			

Automatic annotation of speech signal. The most valuable characteristic of a speech corpus for researchers is a comprehensive annotation of speech signals. The process of manual producing of high quality linguistic annotations is time consuming and requires much effort and linguistic expertise. Phoneme set of speech corpus LRN0 (and later versions) was also employed for automatic annotation of speech on the word level (Laurinčiukaitė et al., 2009). Automatic phonetic labelling of segments is used to create time-aligned annotation files on word or phoneme levels. In the analysed speech recognition engine HTK tools is basically used to train hidden Markov models (HMMs) and to realign data in the recognition process. However, a human annotator must later verify these automatically generated annotations. HTK tools have limited options for a comprehensive investigation of speech data in manual setting of boundaries, therefore additional tool Praat is employed to generate final annotations of speech signals. In Praat the annotation data are stored as a TextGrid annotation file. An example of the Praat TextGrid annotation file for Lithuanian word „akti“ (*eng. To become blind*) is shown in Fig. 2.5.

```
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"
```

Fig. 2.5. Phonetic transcription of the word “akti” from the Praat annotation file

The TextGrid object consists of a number of interval tiers, where each interval tier represents a connected sequence of labelled intervals with boundaries in-between. Interval tiers can represent phoneme-level, phonetic units or word-level transcription. The time boundaries of labelled intervals in the Praat annotation file are measured in seconds.

Manual correction of anotations. The manual correction of Praat annotation files can be done in the following manner. The investigation of boundaries of each word starts with listening to words and visual inspection of their boundaries in a spectrogram and a pitch contour. The pitch contour helps to determine inexact boundaries of the word. Fig. 2.6 shows the main window of Praat where corrections can be made.

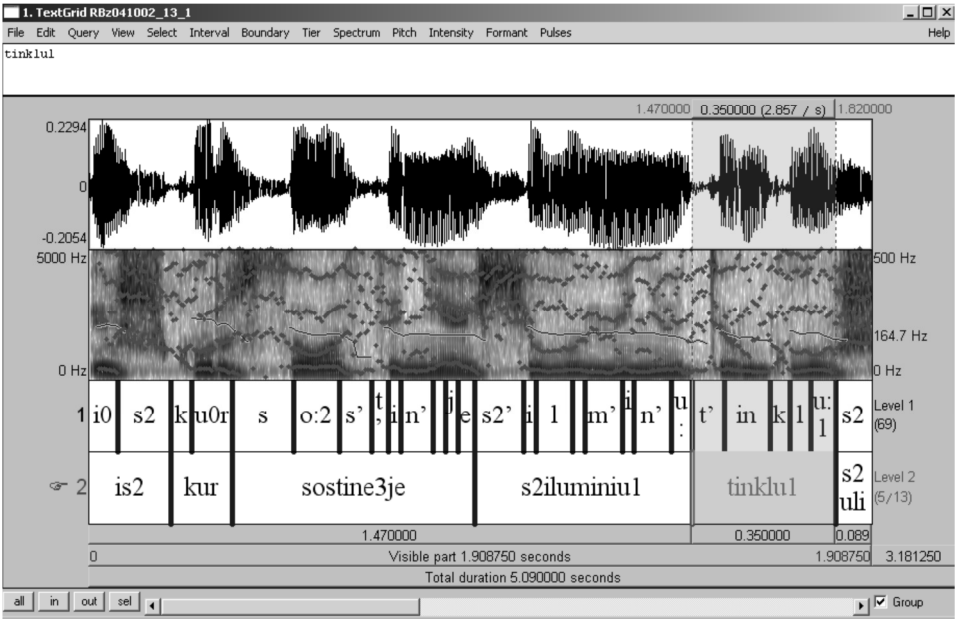


Fig. 2.6. Main window of Praat. Speech waveform and spectrogram (with a pitch contour) of the same sentence with annotations on phoneme-level and word-level are shown

The main patterns of manual correction of the word-level annotations of speech data were as follows (Laurinčiukaitė et al., 2009):

1. Inserting of shorter pauses marked by special conventional words *_pauze* between two vowels.

2. Inexact boundaries of words starting or ending with the consonant r. The corrections were made broadening the segment of the phoneme *r.*
3. Inexact boundary of a word beginning with *im* or *ju.* The corrections were made by broadening the segment of the above-mentioned phoneme combinations.
4. Disappearance of the consonant *s* in collision with consonants *š* and *ž.*
5. Inexact boundary of the word that follows the word ending with s or *š.* The segments of phonemes s, š are too broad and capture a segment of the following word.
6. Inexact boundary of words that collide with two vowels.

Manual correction helps to improve the correctness of automatically generated annotations. However, since manual modification is time consuming and small amount of the corrections are made, Laurinčiukaitė et al. (2009) applied these corrections only to a small segment of the speech corpus. Moreover, the manual revision of automatic speech recognition results was made only in the word-level. Accuracy of phoneme-level annotations was not investigated at all. We need to draw attention to this fact, because it can cause the decrease in speech animation evaluation results (speech animation is strongly related with detected phonemes and their appearance times).

2.3. Translingual visual speech synthesis

During the last years, various "Talking heads" have been proposed, pursuing either a natural 3D facial mesh or video-realistic effects. Most of the existing models are dedicated to animate English language, although recently we also see "Talking heads" driven by Finnish (Olives et al., 1999), Italian (Pelachaud et al., 2001), Chinese Mandarin, Cantonese (Wang et al., 2004) etc.

The speech recognition and animation engine is a critical part of any speech animation system. Building a speech recognition system is data intensive and very tedious and time-consuming task (Xie et al., 2006). So it is very important to explore the possibility to animate a talking face designed based on phonetics in one language, with input audio speech in another (target) language. For example, Verma et al. (Wang et al., 2004) have proposed a Hindi talking face based on a translingual mapping between Hindi and English phonemes and Xie et al. (2006) proposed how an English „Talking head" (Verma et al., 2003) may be driven by input Cantonese speech. Such translingual audio-visual associations enhance the interoperability between audio speech analysis and visual speech

synthesis. Speech recognition engine that understands the phonetic baseforms of the text is required for this.

Some basic aspects must be considered to analyse translingual speech synthesis (Faruquie et al., 2000). Language used to train speech recognition system is called the base language and the language in which the video has to be synthesized is called the novel 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 phoneme set.

Because the aim of mapping the phonemes 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.

Fig. 2.7 (Faruquie et al., 2000) shows the block diagram of the modification layers described above to achieve translingual visual speech synthesis. The subscripts B and N refer to the base language and the novel language respectively and the superscripts P and V refer to phonemes and visemes respectively. The speech recognition system is modified to generate visemic alignments corresponding to the novel language using the phonetic and visemic vocabulary modifiers. In case the visemes for the novel language are available the visemic vocabulary modifier is not required and a direct phoneme to viseme mapping in the novel language may be used to give visemic alignments.

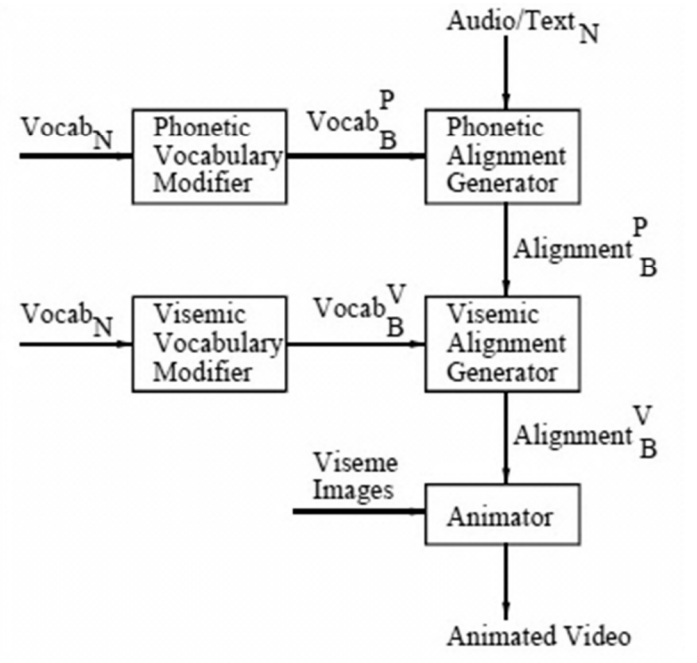


Fig. 2.7. Block diagram showing the modification layers of translingual speech synthesis

Translingual visual speech synthesis is especially useful if the viseme set of the novel language is very different from the viseme set of the base language. In this thesis Lithuanian is the novel language and English is the base language. However, neither Lithuanian visemic vocabulary, neither Visemic vocabulary modifier between Lithuanian and English languages are not defined in the research area yet. Hence, one of the main aims of this thesis is to create translingual visual speech synthesis module and incorporate it into our proposed framework suitable to animate Lithuanian speech. Translingual mapping between phonetic and visemic vocabularies of the analysed languages should be also explored.

2.4. Proposed architecture to animate Lithuanian speech

Many authors of existing „Talking head" software claim that their models are speech independent, but practically audio driven facial animation requires training of a speech visualization system which is used to generate phoneme and

viseme alignments from the input speech. Every language is unique and has specific phonetic rules for speech production. Moreover, the coarticulation effect has a high influence on visual speech also. Modelling of the „Talking head" and construction of the new speech visualization speech is expensive and time consuming processes, so we offer translingual visual speech synthesis as the main concept for the framework to animate Lithuanian speech file.

Framework components are going to be integrated in free/open source facial animation system „iFACE". 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 should be explored. Lithuanian speech recognition engine, translingual visemes selection model, 3D Lithuanian visemes defined, coarticulation modelling and 3D hierarchical head model were merged to get video of Lithuanian „Talking head" synchronized with speech. The overall architecture of our proposed framework is presented in Fig. 2.8.

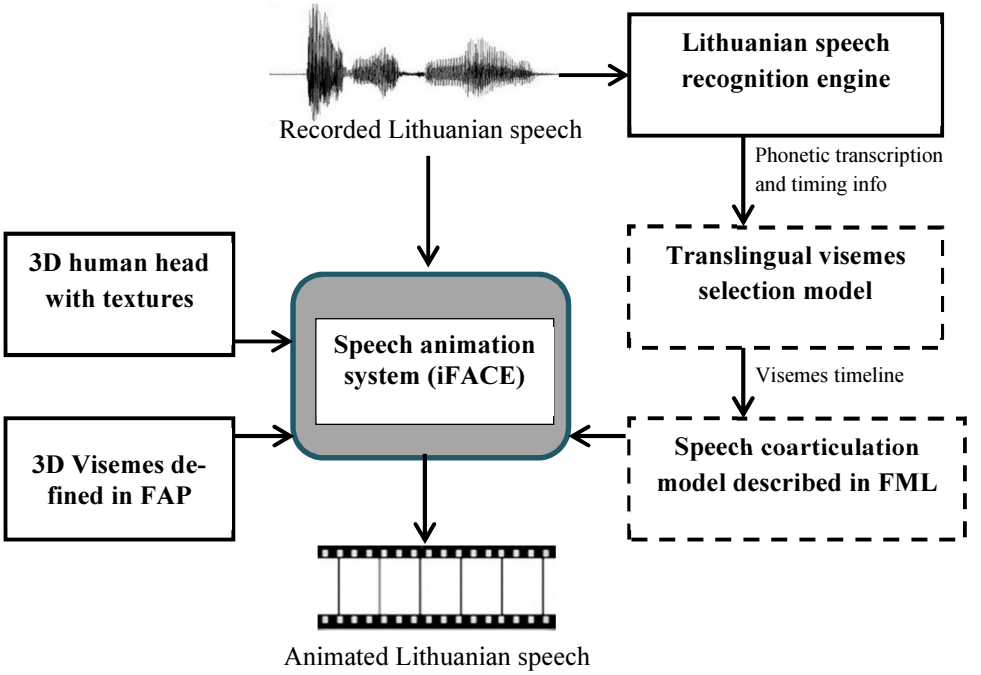


Fig. 2.8. Architecture of the proposed Lithuanian speech animation framework

The data flows are organized as follows:

1. Firstly, phonetic transcription and the timeline of the Lithuanian phonemes are generated by Lithuanian speech recognition engine. The suitable speech recognition engine should be used to identify phonemes in the speech file (*.wav format). Lithuanian speech recognition engine, described in chapter 2.3 was employed in our research.
2. The primary timeline of visemes aligned with phonemes is generated after the new translingual visemes selection model is applied. This model is proposed in chapter 3. This model consists of two main parts: creation of translingual phoneme to viseme mapping table for the most common consonants in the world and the generation of phonemes to visemes mapping table for Lithuanian phonemes that were not defined there.
3. Position of the accented phoneme strongly influences appearance of neighbouring phonetic segments. Besides, visualization of diphthongs requires additional attention, since they regularly appear in Lithuanian words. Coarticulation effect concerns these issues and slightly changes the timeline of visemes. Corrections of the values of expressiveness coefficients of visemes together with their duration information are executed and defined in a single FML file. This file is used to generate the final visualization of the recorded speech. Coarticulation model used in this framework is presented in chapter 3.
4. Lithuanian speech audio file (.wav), 3D geometry head file (.msh) (editable in geometry and texture), the modelled visemes and FML file are input for „iFACE" speech animation system in order to generate animated Lithuanian speech.

The synchronization of speech acoustic and visual output is defined in FML script. Visemes in FML file is described by MPEG-4 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. The next ones deal with specific regions on the face, as left eyebrow, right corner lip, tongue tip, etc. 16 visemes originally described in „iFACE" system (for English language) was edited and used for Lithuanian language. Expressiveness of the visemes was defined after coarticulation control model presented in Section 5 was applied. Example of FML file for animation of Lithuanian word „Ąžuolas" is presented in Fig. 2.9.

```
<?xml version="1.0"?>
<fml>
  <model>
  </model>
  <story>
    <act>
      <param type="FAP" name="1-1-0" value="50" end="200" />
      <param type="FAP" name="1-1-10" value="75" end="260" />
      <param type="FAP" name="1-1-6" value="80" end="94" />
      <param type="FAP" name="1-1-14" value="50" end="84" />
      <param type="FAP" name="1-1-15" value="40" end="84" />
      <param type="FAP" name="1-1-8" value="80" end="68" />
      <param type="FAP" name="1-1-16" value="80" end="64" />
      <param type="FAP" name="1-1-7" value="80" end="113" />
      <param type="FAP" name="1-1-0" value="50" end="400" />
    </act>
  </story>
</fml>
```

Fig. 2.9. FML script for animation of word „Ąžuolas" (translation – „The oak")

So the code line: "<param type="FAP" name="1-1-10" value="75" end="260" />" means, that viseme number 10, has the expressiveness coefficient equal to 75 and is shown for 260 ms.

Separate application to automatically retrieve timing and phonetic information from *.TextGrid file, to apply translingual phoneme-to-viseme mapping module, to integrate coarticulation rules proposed in this paper and to generate the final output as FML file, was written in C++ language.

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

2.5. Conclusions on Chapter 2

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. 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 behavioural 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 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.

However, naturalness of animation in speech-driven „Talking head" strongly depends on language phonetics. „iFACE" was originally created to animate English language, so it doesn't produce satisfactory results when we're trying to get automatically generated syllable transcription and timing information of the recorded Lithuanian speech. In this chapter we presented new framework, suitable to animate Lithuanian speech. Specifics of Lithuanian phonetics, the new translingual speech animation module and the new coarticulation control model are included into proposed architecture.

3

Proposed Lithuanian speech animation technology

Chapter 3 describes proposed Lithuanian speech animation technology which is based on two new models that help to improve intelligibility of animated Lithuanian speech. Translingual visemes selection model divides viseme classes of any language into two parts: constant (translingual) and variable. Lithuanian Coarticulation model is proposed to animate Lithuanian words with pure diphthongs.

The ideas presented in this chapter were published in (Mažonavičiūtė, Baušys 2009; Mažonavičiūtė, Baušys 2010; Mažonavičiūtė, Baušys 2011).

3.1. Translingual visemes selection model

3.1.1. Determination of constant part of viseme classes

Phonemes classification into viseme classes is one of the base problems in speech animation due to the fact that selection of appropriate video segments is based on visemically transcribed phonemes. Human eye cannot visually differentiate between pronunciation of different sounds like /p/ and /b/, while both of them are plosives and look analogous during the speech. So, visually analogous phonemes can be compound into a single viseme class that is illustrated by specific lip shape. Phoneme to viseme mapping remains the key subject in bimodal speech processing due to the fact that correctness of identified visemes considerably influence the realism of animated speech.

Despite the fact, that various researchers attempted to classify phonemes according to their visual similarity, as of date, neither a single classification nor a standardized viseme set exist for individual languages, as it exists for phonemes (Damien et al., 2009). Variability on the number of English viseme classes confirms this assertion: Massaro described 17 viseme groups (Massaro, 1998), meanwhile Dodd and Cambell (Dodd et al., 1984) as well as Tekalp and Ostermann (Tekalp et al., 2000) identified 14 viseme classes.

For this reason audio to visual mapping should be revised for every new language, although it is expensive and time consuming process. We offer to resolve this problem by exploring the idea that groups of internationally common phonemes are visually similar across separate languages. Detailed examination of similarities and distinctions between phoneme-to-viseme mapping tables in different languages is required to identify translingual consonants-to-visemes mapping table.

The general methodology to extract visemes from data of recorded experiments can be described as (Damien et al., 2009):

1. Choice of features, that will be used to observe variation of lip shapes.
2. Definition of word sequences.
3. Recording of the experiment, where native speaker is pronouncing earlier defined phrases.
4. Data processing:

a) normalization, that is necessary to bring the same scale to all distances and areas measured. Also, this operation allows correcting errors, produced by the accidental head movements of the speaker in the camera frame;

b) location of the region of interest. This step reduces the size of the area of segmentation and calculation together with the processing time of each video clip;

c) extraction of characteristic points;

d) calculation of geometric parameters from extracted points.

5. Classification of visemes.

Different features can be selected to explore geometrical transformation of lips while man is talking and to identify separate classes of visually similar phonemes. It causes different quantity of viseme classes in separate languages. Moreover, same phonemes belong to different viseme groups in various languages. Comprehensive analysis of viseme classes in different language is necessary if we want to observe similarities and differences among them.

Phonemes are based on a spoken language and they may range from familiar sounds like [t], [s], [m] to very unusual ones produced in extraordinary ways. The most common vowel system consists of five vowels /i/, /e/, /a/, /o/, /u/ and the most common consonants are /p/, /t/, /k/, /m/, /n/.

Classification of 22 worldly common phonemes into their viseme groups in 11 languages is summed up in Table 3.1. Phonemes are notated by language independent X-SAMPA standard.

Considering the variety of phonemes in different languages, in this section we analyse the set P22 of 22 internationally common consonants:

P22 ={/p/, /t/, /m/, /n/, /b/, /r/, /l/, /f/, /v/, /s/, /z/, /k/, /g/, /d/, /tS/, /dZ/, /S/, /Z/, /h/, /j/, /ts/, /th/}.

We examined consonants of 11 languages to define translingual consonants-to-viseme mapping table for phonemes defined in P22.

L11 = {English, Urdu, Standard Malay, Chinese Mandarin, European Portuguese, French, Mexican-Spanish, Arabic, German, Italian, Lithuanian}

Examination of this table approves conclusion that there is no single agreement on visemes classification. Appropriate clustering method must be applied to define phonemes that are most visually similar across different languages.

Table 3.1. Consonants and their matching viseme groups in 11 languages.

	English	Lithuanian	Urdu	Italian	French	Standard Malay	German	Arabic	Mexican – Spanish	European Portuguese	Chinese Mandarin
l, r	k, g, n, l, h, N, j	t, l, d	t, d, n, l	l	l	n, l	n, l, @n, @l	l	l	l, R, r	l
	w, r	r	w, r	k, g, n, r	R	rl`	h, r, x, N	r	r		Z, tS, S, r
p, b, m	b, p, m	b, m, p	p, b, m	p, b, m	p, b, m	p, b, m	p, b	b m	b, m, p, v	p, b, m	b, p, m
							m				
t, d, k, g, n	t, d, s, z, T, D	t, l, d	t, d, n, l	t, d, s, z, ts, dz	t, d, n	t, d	t, d, k, g	t d z s k n	d, s, t, z	t, d, n	d, t, n
	k, g, n, l, h, N, j	g, k	k, g, N	k, g, n, r	k,g	k,g			c, k, q	k, g, L, J	g, k, h
		n	t, d, n, l		t, d, n	n, l	n, l, @n, @l		n, J	t, d, n	d, t, n
s, z	t, d, s, z, T, D	ts, s, z	s, z	t, d, s, z, ts, dz	s, z	s, z	s, z	t d z s k n	d, s, t, z	s, z	z, ts, s
f, v	f, v	f,v	f, v	f, v	f, v	f, v	f, v	f	-	f, v	f
h	k, g, n, l, h, N, j	-	-	-	-	-	h, r, x, N	h, x, X	-	-	g, k, h
j		j	-	-	j	-	j, C	j	j, g	j	j, q, x
S, Z, tS, dZ	tS , S, Z	tS, S, Z	S, Z, dZ, tS	S, tS, dZ	S, Z	S, Z	S, Z, tS, dZ	Z S	tS, L, y, x	S, Z	Z, tS, S, r
T	t, d, s, z, T, D	-	T, D	-	-	T, D	-	T, D	-		-

Interlanguage confusion matrix for visual consonants

There is a high visual confusion between phonemes in the same viseme class (interclass confusion). However, two visually similar phonemes can be grouped either in a single or in two separate viseme groups in different languages (interlanguage confusion). For instance, pair of phonemes /t/ and /n/ shares the same lip shape in French, but they are splitted in Standard Malay. This aspect must be carefully considered if we want to identify translingual viseme set, which would be suitable to animate languages that are not described in Table 3.1.

The primary task of this research is to decide, which phonemes should be merged in a cluster that describes translingual viseme class. For this reason, the most related elements must be found and merged according to their similarity distance. In this research similarity distance is defined as visual confusion between pairs of consonants influenced by their interlanguage confusion.

The starting point of hierarchical clustering is the construction of similarity matrix. It is a square symmetrical NxN matrix where number in the ith row and the jth column depicts similarity distance between ith and jth elements.

Suppose we have a sequence $P_N = \{p_1, \ldots, p_N\}$, where P_N is set of N analyzed phonemes. Phoneme-to-viseme mappings tables of nr languages defined in set $L_{nr} = \{ l_1, \ldots, l_{nr}\}$ must be investigated to evaluate the interlanguage confusion of visual phonemes. 22 phonemes (P_N) and 11 languages (L_{11}) analyzed in this paper are defined in Table 3.1.

This table reveals that same phonemes belong to different viseme groups. The variable $V_k(l_r)$ is used to characterize the k-th viseme group in the language l_r excluded from data set L_{11}. For instance, $V_k(l_2)$ –is the k-th viseme class in Urdu phoneme-to-viseme mapping table. Similarity distance M(i,j) is used to decide how frequently both phonemes $p_i \in P_N$ and $p_j \in P_N$, where $i, j = 1..N$ belong to the same viseme group across languages. It can be defined as:

$$M(i,j) = \frac{\sum_{l_r=1..nr} S(p_i,p_j)}{nr} \cdot 100\% \quad (1)$$

where $S(p_i, p_j)$ is visual similarity between two phonemes $p_i \in P_N$ and $p_j \in P_N$ in single language $l_r \in L_{nr}$:

$$S(p_i, p_j) = \begin{cases} 1, & \text{if } (p_i \text{ and } p_j) \in V_k(l_r) \\ 0, & \text{if } (p_i \text{ and } p_j) \notin V_k(l_r) \end{cases} \quad (2)$$

Interlanguage confusion matrix for consonants is filled with values $M(i,j)$, which describes similarity distance between phonemes $p_i \in P_N$ and $p_j \in P_N$.

To improve visualization of results interlanguage confusion matrix, where darker zones mean higher visual correlation between phonemes in the corresponding row and column is demonstrated in Fig. 3.1.

Visual similarity between pairs of consonants was measured independently from the others: no item of data contributes to more than one distance. The diagonal elements are not considered, since the confusion between an object and itself is postulated as 100%. The greater similarity percent among phonemes is identified, the greater is possibility that two phonemes belong to the same viseme class.

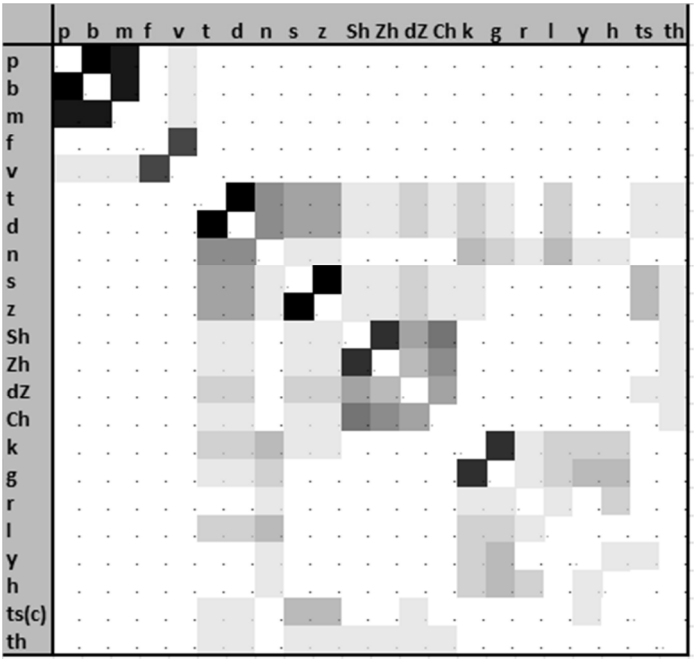


Fig. 3.1. Interlanguage confusion matrix for consonant, where darker zones means higher visual correlation between phonemes in the ith row and the jth column

There is clearly demonstrated, that 8 phoneme duets: {/p/, /b/}, {/b/, /m/}, {/p/, /m/}, {/f/, /v/}, {/t/, /d/}, {/s/, /z/}, {/S/, /Z/} and {/k/, /g/} always appear in the same cluster if they both are defined in specific language. However, interlanguage confusion matrix defines consonants, which are most alike pairwise and it doesn't take into consideration the fact, that there are additional phonemes that belong to that clusters too. For example, Italian phonemes /k/, /g/, /n/, /r/ are considered as visually similar and they are considered as single

viseme class, while in Chinese Mandarin situation is diverted – consonants /k/, /g/ are grouped together with phoneme /h/.

Distinct viseme class can be composed either from a single, or from two and more phonemes (Table 3.1). Additional mechanism must be applied to identify the final number of interlanguage viseme classes and visual consonants that should be congregated in them.

Hierarchical clustering of visually similar consonants

Evaluation of all pair-wise similarity distances between visual phonemes and construction of their confusion matrix are the primary steps for data analysis using hierarchical clustering method. This technique does not require a prespecified number of final clusters, so it is highly suitable for our research since the number of viseme classes in separate languages are different and optimal count still is not identified in the world.

We propose to utilize Hierarchical agglomerative clustering (HAC) to identify clusters for N consonants having the NxN interlanguage confusion matrix. The methodology can be described as:

1) Start with each phoneme $p_i \in P_N$ assigned to its own cluster C_j $(i, j = 1..N)$.
2) Until all of the data are agglomerated in one cluster:
    a) Determine the two clusters $\mathrm{X} = \mathrm{C_j}$ and $\mathrm{Y} = \mathrm{C_j}$, that are most similar among the current clusters.
    b) Remove the pair $\mathrm{C_j}$ and $\mathrm{C_j}$ from the interlanguage confusion matrix $\mathrm{M(i,j)}$ and update the matrix by appending the new row and column corresponding to newly formed cluster $t = \mathrm{C_i} \cup \mathrm{C_j} = \boldsymbol{sim}(\mathrm{X,Y})$.
    c) Similarity distances $\mathrm{sim(t,z)}$ between the new cluster t and each of the old clusters $\mathrm{z} \neq \mathrm{X,Y}$ must be computed and matrix $\mathrm{M(i,j)}$ should be updated.

The function that determines the similarity of two clusters is described as sim(x,y). In the first iteration similarity between the clusters equal the similarity distances between the phonemes they contain and indexes i and j are described as consonant numbers in the phoneme sequence $\mathrm{P_N}$.

From the second iteration each of two clusters possibly contains multiple phonemes. So, we have many choices to define intergroup similarity of distance measures between analysed objects. Three most popular methods are:

1. Single Link clustering that is based on similarity of two most similar members.
2. Complete Link clustering works with similarity of two least similar members.

3. Group Avarage clustering is compromise between single and complete link. It uses average similarity across all ordered pairs of objects within the merged cluster to measure the similarity of two clusters, so it avoids the pitfalls of the previous methods, when similarity is equated based on a single pair of analysed objects.

We propose to use Group-average agglomerative clustering (GAAC) method to define similarity $\mathrm{sim}(X, Y)$ between two clusters X and Y. Mathematically it can be described by the following expression:

$$\mathrm{sim}(X, Y) = \frac{1}{N_X N_Y} \sum_{i=1}^{N_X} \sum_{j=1}^{N_Y} d_{ij} \tag{3}$$

where N_X and N_Y are the count of consonants in clusters X and Y; d_{ij} is the similarity distance between elements in the ith row and the jth column of the current similarity matrix $M(i, j)$.

The result of HAC clustering is typically visualized as a dendrogram, where each merge is represented by a horizontal line.

Extraction of viseme classes by dendrogram analysis

As discussed earlier, Hierarchical Group-average agglomerative clustering method starts with single phoneme in a single cluster and later, the closest pair of clusters is iteratively merged until all of the data are agglomerated in one group. The whole clustering process of N analysed phonemes from the level when each cluster contains one consonant up to the level where a single cluster contains all analysed consonants is represented as dendrogram in Fig. 3.2.

The ∏-shaped lines that connect objects in a binary tree in the dendragram are inspected to observe classification results of visual consonants. HAC does not require a prespecified number of clusters and it is up to the researcher to define criterias that should be used to determine the cutting point of the dendrogram that influence the count of viseme groups.

Length of the ∏-branch is directly proportional to the distinctness between analyzed clusters and the height of each ∏-branch represents inversely proportional relation between connected clusters, thus short vertical lines indicate larger visual similarity between groups of visemes. For example, the vertical height of the branch connecting phonemes /k/ and /g/ is 18%, meanwhile height between cluster {/k/, /g/} and viseme /j/ is 40%. It supposes that phonemes /k/ and /g/ are more visually similar then duet /k/ and /j/. Cautious analysis of the dendragram is necessary to come to a decision what is the final count and content of inter-language viseme classes.

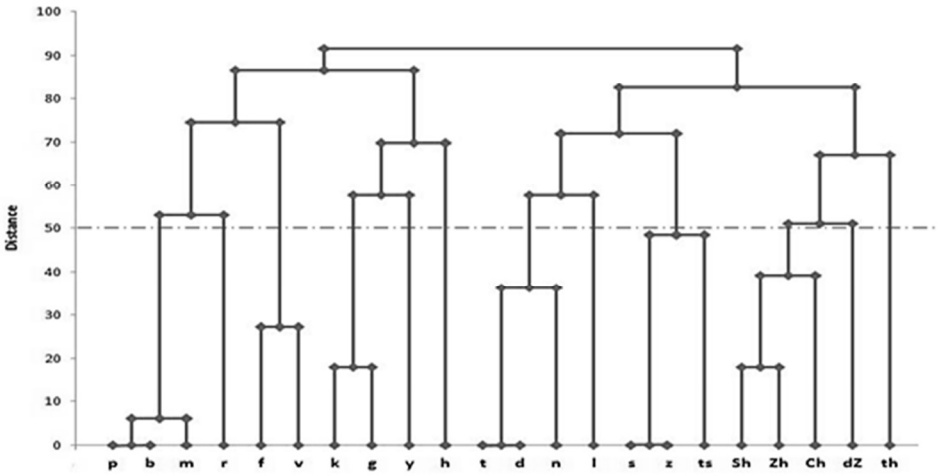


Fig. 3.2. Dendragram that illustrates consonants clustering into viseme classes with reference to visual similarity of phonemes, that were defined in viseme classification tables in 11 languages. Group-average agglomerative clustering method was used to define similarity between two clusters

Different criteria can be used to determine the cutting line:

a) Cut the dendrogram where the gap between two successive combination similarities is largest. Such large gaps arguably indicate „natural" clustering.

b) Cut at a prespecified level of similarity.

There is no explicit gap between clusters in dendrogram in Fig. 3.2, so we have decided to utilize the second option for our task. Hierarchical clustering should be treated with caution, because improper cutting line influences erroneous results of the research.

Careful examination of the dendrogram and its relation to the analysed situation should be performed to achieve high-quality clustering. Taking into account the existing classifications of viseme classes in different languages (Table 3.1.), we defined that the distance that breaks dendragram into two parts is equal to 52. This vertical cutting line is marked in Fig. 3.2. Consequently, if the horizontal branch of the cluster is below this line, all visemes of the cluster can be classified as single viseme class, otherwise, if horizontal ∏-branch is above line, we don't analyse it.

Final conclusions on translingual visemes clustering by dendragram analysis are demonstrated in translingual consonants-to-visemes mapping table (Fig.

3.3). 11 interlanguage viseme classes for 22 consonants based on phoneme-to-viseme mappings in 11 different languages were identified.

Viseme class	Visual representation	Viseme class	Visual representation
p, b, m		t,d,n	
r		l	
f,v		s, z, ts	
j		Sh, Zh, Ch, dZ	
h		T	
k, g			

Fig. 3.3. Consonant viseme groups calculated by using Group Avarage clustering method for 22 internationally common consonants

According to the results, we propose 7 recommendations for viseme classification:

1) Define separate visemes for phonemes /r/, /j/, /h/ and /l/. Interlanguage confusion matrix (Fig. 3.1.) demonstrates that consonants /r/, /j/, /h/ and /l/ have a pure visual correlation with other phonemes:

- Phoneme /r/ with /k/, /g/, /l/, /h/;
- Phoneme /j/ with /n/, /k/, /g/, /h/, /ts/;
- Phoneme /h/ with /k/, /g/, /r/, /y/;
- Phoneme /l/ with /t/, /d/, /n/, /k/, /g/ and /r/.

Similarity distances between each of /r/, /j/, /h/, /l/ and other consonants are less than 30% (Fig. 3.2.). On the basis of proposed GAAC clustering, we recommend to define 4 viseme classes with a single consonant (/r/, /j/, /h/ or /l/) in each.

2) Identify distinct viseme for phonemes /T/ and /D/. Phoneme /T/ is defined in a few analysed languages; however it is enough to make a decision which group of visemes it belongs to. Similarity between /T/ and other analysed consonants is less then 10%, so /T/ must be separated from them. Phoneme /D/ was not included in this research, however, consonants /T/ and /D/ always appear in the same viseme class (Table 3.1.), so we recommend to define a distinct viseme class for /T/ and /D/.
3) Add phoneme /n/ in one group with /t/ and /d/. There is a strong disagreement on visual similarity between /n/ and other consonants across languages. It is grouped with /k/, /g/, /l/, /h/, /y/, /t/, /d/, /r/, /s/, /z/ in different phoneme-to-viseme mapping tables; however dendragram of hierarchical clustering eliminates such confusion and shows a high interlanguage visual correlation between /t/, /d/ and /n/.
4) Assembly phonemes /S/, /Z/ and /dZ/ into a single viseme class. Nevertheless, that consonant/dZ/is defined in 3 from 11 analysed languages and consonant /tS/is included in 7 of them, dendragram analysis suggests defining of/dZ/ and /tS/ together with /S/ and /Z/ in a single viseme class.
5) Add phoneme /ts/ into single viseme class with /s/ and /z/ consonants. On the basis of proposed GAAC clustering, we suggest to define /ts/ as visually similar with /s/ and /z/, even though it was defined just in 3 phoneme-to-viseme mapping tables internationally (Table 3.1.).
6) Exclude duet of phonemes /k/ and /g/ as the separate viseme class. The pair of visual phonemes /k/, /g/ has a visual confusion of 82%. It is proposed to group them with phonemes /n/, /l/, /h/, /j/, /r/, /t/or/d/across languages, but interlanguage visual similarity between /k/ or /g/ and other consonants are less than 30%. Dendragram approves this analysis and suggests grouping of /k/ and /g/ as a distinct viseme class.
7) Exclude two groups of consonants {/p/, /b/, /m/} and {/f/, /v/} as different viseme classes. Similarity distance between phonemes /p/, /b/, /m/ is higher then 91 % and it is equal to 73% among consonants /f/ and /v/, so two viseme classes can be evidently distinguished.

Translingual evaluation of clustering results

11 interlanguage viseme classes based on investigation of visual confusion of consonants in 11 languages are the result of our research. However, additional study is required to evaluate correlation between our proposed classification of visemes and phoneme-to-viseme mappings defined in separate languages.

Accuracy of our research can be evaluated by compliance between interlanguage GAAC clustering of visual consonants (Fig. 3.4) and viseme groups, identified in different languages (Table 3.1.). Suppose that k is the index of viseme class $V_k(I_n)$ in the interlanguage consonants-to-viseme mapping table and m is the indicator of viseme group $V_m(l_r)$ in phoneme-to-viseme mapping table of separate language l_r. r is the index of language in the set of analyzed languages:

L_{11} = {English, Urdu, Standard Malay, Chinese Mandarin, European Portuguese, French, Mexican-Spanish, Arabic, German, Italian, Lithuanian};

We propose to estimate this conformity by following 3 criteria:

1) The compliance $A(V_k(In), V_m(l_r))$ between viseme classes $V_k(In)$ and $V_m(l_r)$ is equal to 1, IF total set of consonants $P_K = \{p_{k1}, .., p_{kk}\}$), which forms interlanguage viseme class $V_k(In)$, coincides with the whole set of phonemes $P_M = \{p_{m1}, .., p_{mm}\}$) in viseme group $V_m(l_r)$
2) ELSE IF complete group of phonemes $P_K = \{p_{k1}, .., p_{kk}\} \in V_k(In)$ is the partial heap of $P_M = \{p_{m1}, .., p_{mm}\} \in V_m(l_r)$, compliance of interlingual viseme class k in language l_r is 2/3. This criteria was chosen to highlight the situation when all elements of $V_k(In)$ are included in $V_m(l_r)$, but the number of phonemes defined in interlingual viseme class k is smaller than the count of consonants in viseme group of specific language l_r. Despite the fact that smaller groups of units in a single viseme class makes speech animation more accurate, we defined the accuracy of this case as 2/3. This choice is based on the fact, that compression of visually similar phonemes is lost when dividing viseme class into smaller groups of visemes.
3) OTHERWISE, compliance $A(V_k(In), V_m(l_r))$ is equal to 0.

Mathematically, it can be described as:

$$A(V_k(In), V_m(l_r)) = \begin{cases} 1, \text{if } (P_K \in V_k(In)) = ((P_M \in V_m(l_r)) \\ \frac{2}{3}, \text{if } (P_K \in V_k(In)) \subset ((P_M \in V_m(l_r)) \\ 0, \text{if } (P_K \in V_k(In)) \not\subset ((P_M \in V_m(l_r)) \end{cases} \quad (4)$$

The compliance $A(V_k(In), V_m(l_r))$ of interlingual viseme group k in language l_r is used to evaluate the precision of similarity $Pr(l_r)$ between phoneme-to-viseme mapping table of interlanguage clustering (Fig. 3.3.) and viseme classes defined in separate language l_r (Table 3.1.). Precision of similarity of mapping tables $Pr(l_r)$ is defined as:

$$Pr(l_r) = \frac{\sum_{l=1..nr} A(V_k(In), V_m(l_r))}{nr} \cdot 100\% \quad (5)$$

where nr is the total count of visemes (in language l_r), which contain single or more of traced phonemes from the consonants set P_N.

Similarity between new interlanguage consonants-to-viseme mapping table and phoneme-to-viseme mappings for different languages, earlier presented by various researchers are summarized in Fig. 3.4.

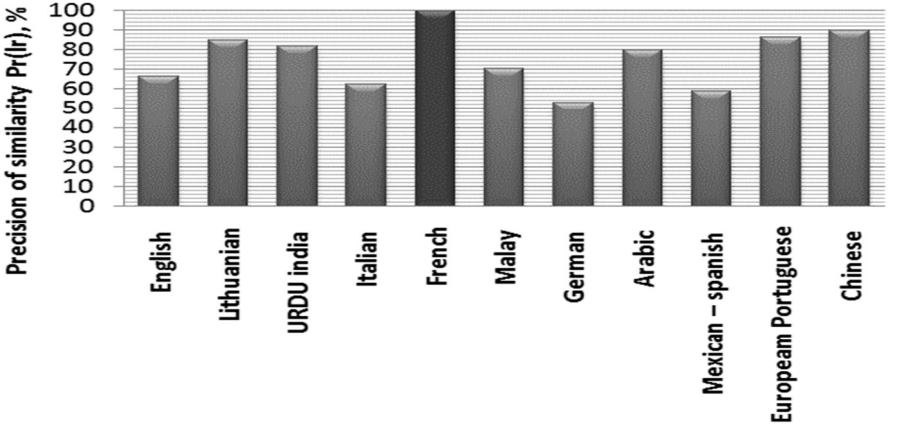


Fig. 3.4. Similarity between new interlanguage consonants-to-viseme mapping table and phoneme-to-viseme mappings for different languages

High correlation between interlanguage viseme classes and viseme groups across languages is revealed in Fig. 3.4. The lowest precision of similarity is defined for German language (54%) and the highest – for French (100%!). However, precision of similarities for the most widely spoken languages (Urdu, French, Arabic and Chinese) is higher or equal to 80%. Moreover, we firmly believe that German phoneme-to-viseme mapping technique, which did not use

objective criteria to identify viseme classes, is credible cause to affect the lowest precision of similarity.

3.1.2. Analysis of variable part of viseme classes

There are two kinds of phonemes in any language: phonemes that frequently reappear in other languages, and those who are specific for analysed language. Accordingly, consonants set of any language can be splitted into two groups: phonemes represented in translingual consonants-to-viseme mapping table and those, which are not described there. For instance, Lithuanian letters like ą, ę, ė, į, ų, ū don't have a corresponding viseme defined there. Additional research must be performed to group the resident Lithuanian phonemes with respect to their visual similarity. „iFACE" speech animation system has to visualize Lithuanian language using viseme set of English language, so visemic alignment cannot be simply generated from the phonetic alignment using direct phoneme to viseme mapping defined in the engine. Lithuanian to English viseme mapping table should be defined to animate specific Lithuanian phonemes.

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. English phoneme set consist of 48 phonemes (this count varies). Lithuanian and English phonemes can be related according to the table proposed by Kasparaitis (2008). Visual similarity between Lithuanian and English visemes should be analysed to create Visemic vocabulary modifier between Lithuanian and English languages.

The English speech recognition and animation system „Crazy Talk" (commercial software) was picked for this task (Fig. 3.5.).

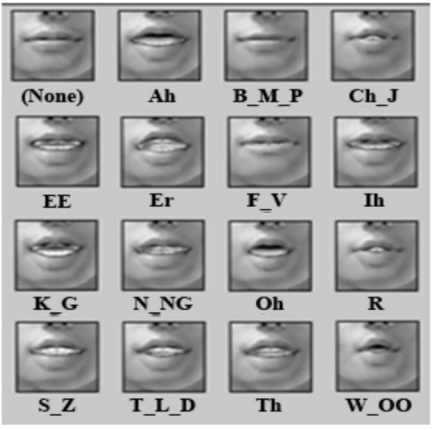


Fig. 3.5. Visemes classification in „Crazy Talk" software

This system offers phoneme-level speech recognition. After some experiments we noticed that phonetic transcription is quite precisely rearranged in the timeline even when Lithuanian speech file is used as input. On the other hand, English phonetic vocabulary is used for speech recognition, so associated visemes are chosen mistakenly for Lithuanian speech. 16 visemes defined in „Crazy Talk" is shown in Fig. 3.5. However, there is possibility to manually replace the wrong English viseme with the right one, if we would know what English viseme must be placed at the detected time of particular Lithuanian sound.

3D Lithuanian visemes should be identified and analysed to group visually similar phonemes into English viseme groups. If Lithuanian phoneme cannot be related with any English viseme class, individual viseme group should be created and separate viseme should be modelled for this phoneme in the chosen speech animation system.

Modelling of 3D Lithuanian visemes

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. 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 Fig. 3.6.

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 C1 and C2, which lie within the physical camera and are at a distance b from each other (Grammalidis et al., 2002); f is the focal length.

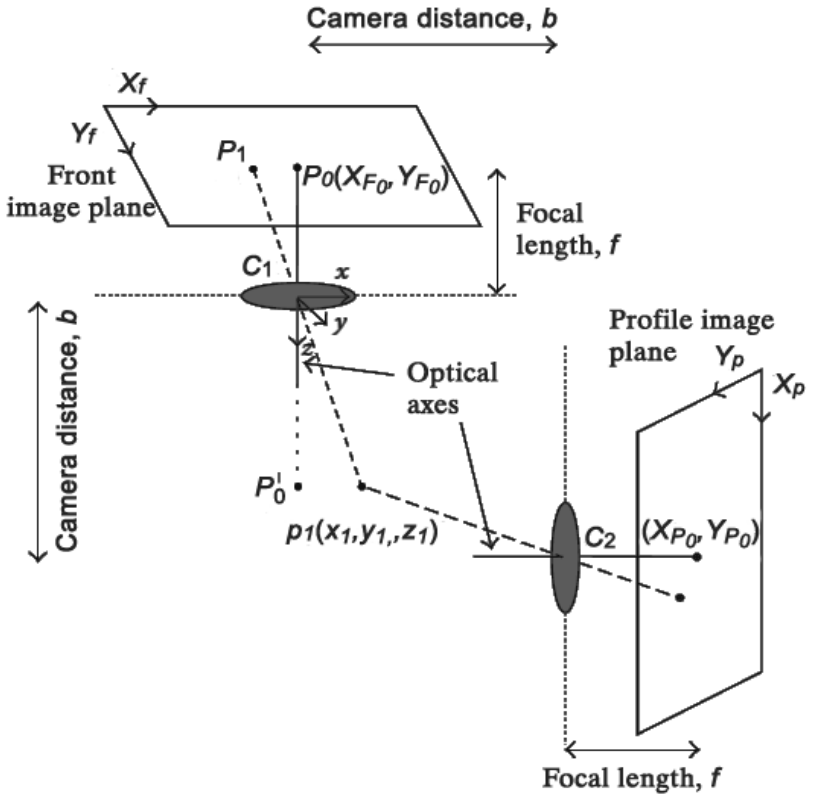


Fig. 3.6. Image acquisition layout for the experiment

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:

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

$$X_P = f\frac{b-z}{b-x} + X_{P_0}, \quad Y_P = f\frac{-y}{b-x} + Y_{P_0,} \qquad (6)$$

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

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.

We have analysed 30 Lithuanian phonemes, transcribed by SAMPA standard (Raskinis et al., 2003). 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 analysed Lithuanian phonemes.

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:

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = R \begin{bmatrix} x_0 \\ y_0 \\ z_0 \end{bmatrix} + T \qquad (7)$$

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 $[T_x \quad T_y \quad T_z]^T$.

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 influences the quality of animation. They are speech recognition engine, naturally looking 3D visemes and the correctness of phoneme to viseme mapping table.

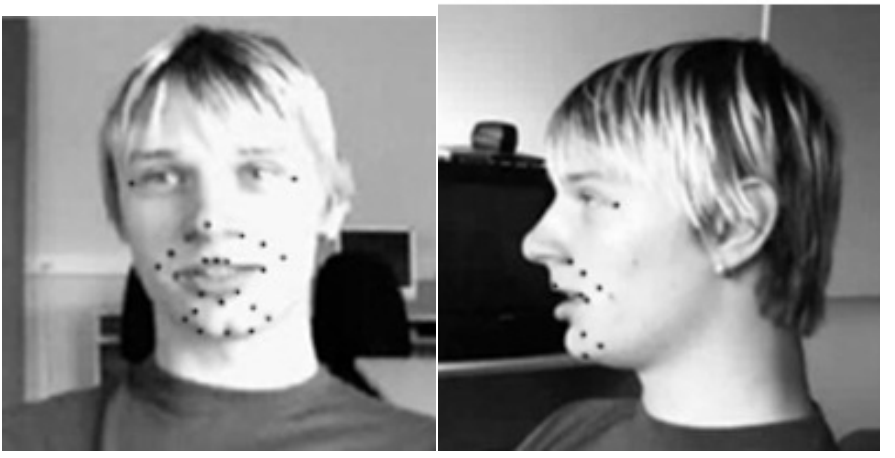
The first one is speech recognition engine and its accurateness when we are trying to detect the time information when the new phoneme appears. In order to find the Lithuanian phoneme to viseme mapping table, the input sound file has to be extracted from video material of speech. The extracted acoustical 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 analyse 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 system 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 looks 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 the Fig. 3.7.

For the creation of Lithuanian 3D viseme we've used the modified technology of static viseme modelling (Pighin et al., 1998). 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.



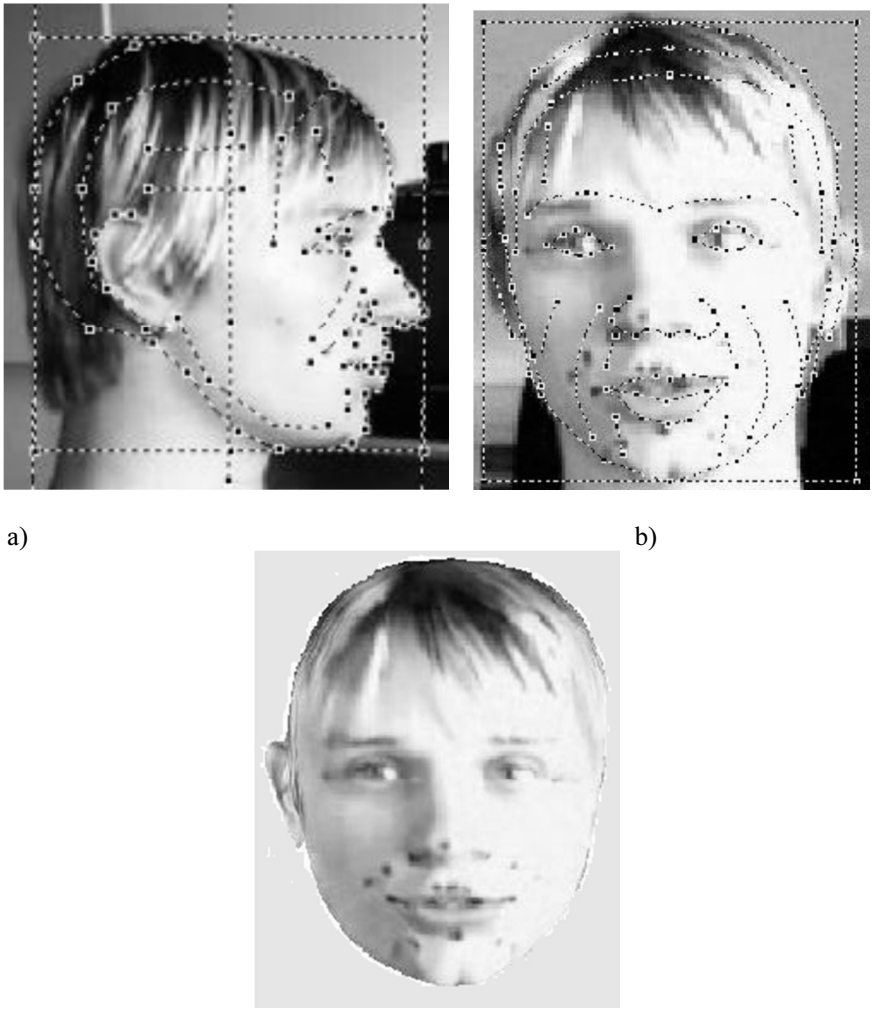
a) b)

Fig. 3.7. 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

To implement the technology, we've used the freeware „Faceworks", 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 Fig. 3.8.

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".



a) b)

c)

Fig. 3.8. 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

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 varies 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. Due to the magnitude of lips feature points moving vectors, the same English viseme can be matched as different Lithuanian visemes. 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 viseme /ʒ/ is shown in Fig. 3.9. As we see, Lithuanian 3D viseme of the phoneme /ʒ/ almost matches English viseme "Ih" (expressiveness 100), but the more accurate result was defined using the English viseme "Ch_J" (expressiveness 80).

Fig. 3.9. 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

According to the results, English and Lithuanian visemes were related in the Lithuanian viseme to English viseme mapping table (Table 3.2.). Parameter of expressiveness and Lithuanian letter representing the corresponding phoneme are defined in the table, too.

Table 3.2. Lithuanian viseme to English viseme mapping table

English viseme	Lithuanian viseme, 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	/ʧ/ 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)

The generated mapping table was successfully applied in the animation software „Talking head" by manually replacing automatically wrongly identified visemes with the right ones. It also works as translingual visemic vocabulary between Lithuanian and English languages.

3.1.3. Application of translingual visemes selection model

Possibility to reuse already existing speech animation system 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. The purpose of this section is to enlighten new aspects of translingual visemes mapping that is included into the speech animation architecture suitable for Lithuanian Speech Animation.

„iFACE" employs English phonetics alignment generator with corresponding viseme set. If it is applied to animate Lithuanian language, direct acoustic to visual linkage (mapping) fails to produce convincing speech visualization. Thus, we must define and employ translingual visemes selection model to animate Lithuanian speech using constant (translingual) and variable sets of viseme classes.

If Lithuanian phoneme belongs to the set of 22 internationally common consonants, it should be visualized using visual representation of translingual viseme classes defined in Fig. 3.3. Otherwise, variable set of viseme classes should be used for audio to visual mapping. Since English is the base language of the system, that is chosen to animate Lithuanian language, additional vocabulary that relates Lithuanian phonemes with Lithuanian and English visemes can be used to define variable part of Lithuanian visemes (Table 3.2.). Additional visemes should be modelled for specific Lithuanian phonemes that do not have corresponding viseme defined in constant or variable sets of viseme classes.

Lithuanian Visemes (*LV*) can be divided into 2 groups: those who are defined in 11 Translingual visemes classes (*TV*) and those who don't. Construction of our proposed translingual visemes selection model used for Lithuanian speech animation is described as:

a) If *LV* has a matching *TV*, constant viseme classes should be applied to find the most suitable visual representation for the analysed Lithuanian consonant.

b) If *LV* has not a matching *TV* and Lithuanian Phoneme (*LP*) has the analogous English Phoneme (*EP*) in *LP-EP* mapping table proposed by Kasparaitis (2008), English Phoneme to English Viseme (*EV*) mapping table basically employed in the chosen speech animation system, should be used to find the most suitable *EV* for the analyzed *LV*. Expressiveness parameter of EV should be altered to manage phonetic features of Lithuanian language.

c) If *LV* has not a matching *TV* or *EV* (e.g. Į, Ė), they can be defined as specific Lithuanian phonemes (*SLV*), that should be sculpted individually to animate Lithuanian speech. New visemes is defined by the points of MPEG-4 standard.

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

There are 4 items that strongly influence the quality of speech animation: accuracy of speech recognition engine, naturally looking 3D visemes, the correctness of phoneme to viseme mapping and speech synchronisation algorithm. Middle

two of them take advantage of the proposed translingual visemes selection model, which is also suitable to animate foreign languages:

1. Application of proposed model significantly reduces the amount of new visemes necessary to be defined and sculpted to animate novel language. Examination of the relation between LV and EV showed that only 17% of 35 Lithuanian phonemes must be analysed for visual representation. Residual Lithuanian phonemes can be represented by Translingual viseme set or English viseme classes.
2. The model offers possibility to reuse visemes of the chosen speech animation system, although small modifications can be required: visual representations of SLP can be appended as separate units.

Translingual visemes selection model is part of the proposed architecture of speech animation framework suitable for Lithuanian Speech Animation.

3.1.4. Concluding remarks on translingual visemes selection model

In this section we defined 11 interlanguage viseme classes and proposed 7 recommendations for consonants clustering into those groups. We have analysed 22 internationally common consonants and presented novel method to define translingual viseme classes. Our work is relevant since phonemes-to-viseme classification noticeably differs across separate languages and neither single standardized set of viseme classes was proposed so far, although the same phonemes were analysed in various investigations. Evaluation of compliance between novel interlanguage consonants-to-visemes mapping table and those, earlier defined by various researches, was also proposed.

Translingual visemes selection model defined in our research can be successfully applied for new speech animation systems, because viseme set of any speech can be split into two groups: constant and variable.

3.2. Lithuanian coarticulation model for pure Lithuanian diphthongs

3.2.1. Coarticulation analysis of Lithuanian diphthongs

The term diphthong refers to two adjacent vowel sounds occurring within the same syllable. Technically, a diphthong is a vowel with two different targets. This means that tongue and lips move during the pronunciation of this vowel.

Languages differ in the length of diphthongs. Diphthongs typically behave like long vowels in languages with phonemically short and long vowels. Lithuanian language is a good example of this case. Besides, languages differ in the count of diphthongs (10 in British English, 6 in Dutch, etc.).

There are two types of Lithuanian diphthongs: 9 pure (*Vowel-Vowel structure (VV))* and 20 mixed diphthongs, that are made of vowels "a", "e", "i", "u" and consonants "l", "m", "n", "r" and has the *Vowel-Consonant structure (VC).* Only pure Lithuanian diphthongs (*ai, au, ei, ui, ie, uo, eu, oi, ou*) will be analysed in our research. They regularly appear in Lithuanian words (e.g. *miegas, saulė, eisena*), so visualization of these diphthongs requires additional attention; especially when English speech animation system is used to animate Lithuanian speech.

Stressing of the diphthong highly influence visualization of the speech since stressed syllable is more expressive then others. Besides, position of the accented phoneme strongly influences appearance of neighbouring phonetic segments. English diphthongs are always stressed with the falling accent and Lithuanian diphthongs can be stressed with rising accent, too. Thus, three situations of Lithuanian diphthong stressing can be distinguished:

a) ***Diphthong is stressed with falling accent (ái).*** Falling diphthong starts with a vowel quality of higher importance (higher pitch or volume) and ends in a semivowel with less prominence. Examples of falling diphthongs: *l**ái**mė, **áu**gti, l**éi**sti.*

b) ***Diphthong is stressed with rising accent*** (aĩ). Rising diphthong begins with a less prominent semivowel and end with a more prominent full vowel. Examples: ***eĩ**ti, šiẽnas.*

c) ***Diphthong is in the unstressed syllable*** (*e.g. tr**au**kinys, ąžu**o**las*).

People prepare themselves for pronunciation of the next phoneme during articulation of the current phoneme. So, the second group of features that influence visualization and expressiveness of Lithuanian diphthongs includes its location in the word and its neighbours.

Three cases of diphthong location in the word can be distinguished:

a) End of word is one letter to the right of the current diphthong (VVC structure) (e.g. "tak**ai**s).

b) End of word is in the diphthong (CVV structure) (e.g. „tak**ai**").

c) End of word is somewhere else (VVCV, VVCCV etc. structures) (e.g. „t**ai**ka, **ai**tvaras").

Diphthong position in the word together with information about its stressing are analysed to investigate their influence for speech animation. Rules for Lithuanian diphthong visualization are included in the framework for Lithuanian diphthong visualization (Fig. 3.10).

3.2.2. Proposed framework for pure Lithuanian diphthong visualization

Rules for pure Lithuanian diphthong visualization (Fig. 3.10) can be divided into two main groups: rules where adapted Ohman coarticulation control model can be applied (dotted and greyed rules) and those, where it cannot be done (white rules).

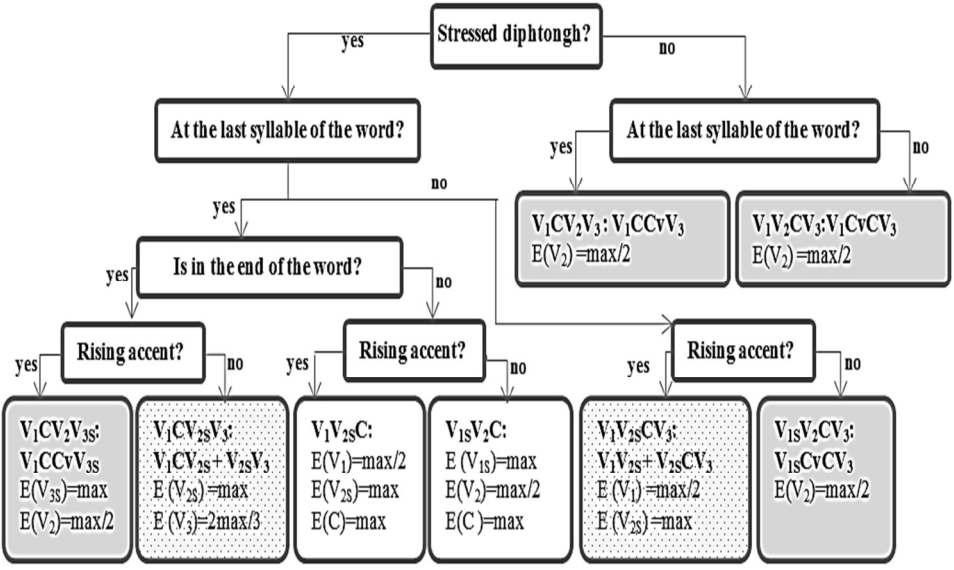


Fig. 3.10. Framework for pure Lithuanian diphthong animation. Dotted and greyed rules define situation when Ohman coarticulation control model can be applied and white rules define state, where it cannot be done

First group of rules uses Vowel-Vowel-Consonant-Vowel (VVCV) or Vowel-Consonant-Vowel-Vowel (VCVV) utterances for diphthong visualization, where structure VV defines a pure Lithuanian diphthong.

Ohman coarticulation control model proposed earlier does not define coarticulation between two consonants. So, the influence of analysed consonant must extend no further than to the peak of the preceding or following gesture. Moreover, Ohman model (Ohman, 1967) is designed for VCV, VCCV or VCCCV

phonetic utterances, therefore its application for VVCV and VCVV utterances must be considered separately. So, in this paper we propose technique, how VVCV and VCVV utterances can be visualized using Ohman coarticulation model designed for VCCV phonetic structure.

Lithuanian diphthong can be stressed in three ways (falling accent, rising accent or non-stressed), so the influence of the stressed vowel V_S for the appearance of non-accented vowel of diphthong should be analysed. It was stated earlier, that falling diphthong ($V_{1S}V_2$) starts with a vowel quality of higher importance and ends in a semivowel with less prominence. It means that semivowel ($V_{semi} = V_2$) is visually much less expressed too. Therefore we define that supreme expressiveness coefficient $E(V_{semi})$ for viseme of semivowel is equal to the half of vowel's V_2 maximum expressiveness:

$$E(V_{semi}) = \frac{E(V_{2\max})}{2} \tag{8}$$

The rising diphthong V_1V_{2S} begins with a semivowel and ends with a more prominent full vowel, so $V_{semi} = V_1$. In the meantime, we treat the non-stressed (V_1V_2) diphthong as duet of two semivowels ($V_1V_2 = V_{1semi}V_{2semi}$).

Vowel articulation strongly influences pronunciation of neighbouring phonemes. Consonants' visual expressiveness is much lower, so they are highly dependable from neighbouring vowels. In the meantime semivowel has similar characteristics as consonant: it is highly influenced by stressed vowel and its expressivity is much lower. So in VVCV and VCVV utterances, we propose to treat semivowel as virtual consonant (C_V), which has the maximum expressiveness values equal to the maximum expressiveness of semivowel. This transcription gives us possibility to use Ohman coarticulation rule model for diphthongs in VCVV and VCVV syllables. For instance: syllable $V_{1S}V_2CV_3$ can be transformed into utterance $V_1C_VCV_3$, which is suitable for diphthong visualization with Ohman model.

The vowel track $v(t)$ is formed by interpolation between fully expressed vowel targets (visemes). So, this model can be applicable to track parameters of the visemes that appear in phonetic structures V_1CV_2, V_1CCV_2, V_1CCCV_2 etc. A consonant is specified by a target value c, a coarticulation factor w_c and a function $k(t)$ that dictates the temporal blending of vowel track and consonant target. Both w_c and $k(t)$ are in the interval [0, 1]. The trajectory of a given articulatory parameter over arbitrary phoneme sequences can be described as:

$$z(t) = v(t) + \sum_{i \in C} w_{ci} k_i(t)(c_i - v(t)) \tag{9}$$

where C is the set of all consonants in the analyzed phonetic structure.

The vowel track $v(t)$ is formed by temporally blending successive fixed vowel targets a_i, according to the function:

$$v(t) = \frac{\sum_{j=1}^{N} a_j b_j(t)}{\sum_{j=1}^{N} b_j(t)} \tag{10}$$

where N is the number of vowels in the analysed utterance and $b_j(t)$ is the blend function of the j^{th} vowel in the utterance. Cubic function $b_j(t)$ has the value $b_j(t) = 1$ at the centre of vowel j and the value $b_j(t) = 0$ at the centre of the preceding (j-1) and following (j+1) vowel.

Since there is no intervocalic coarticulation in the Ohman model, the blending function $b_j(t)$ can be also applied to consonants that are between vowels.

Different model should be dapplied for coarticulation rules that are the dotted in Fig. 3.10. Analysis of the phonetic structure $V_1V_{2S}CV_3$ shows, that it should be transformed to $C_VV_{2S}CV_3$ utterance, but the whole phonetic structure is not suitable for Ohman model. On the other hand, structure $V_1V_{2S}CV_3$ can be split into two parts: V_1V_{2S} and $V_{2S}CV_3$ utterances. Coarticulation model can be applied for $V_{2S}CV_3$ utterance and linear interpolation can be applied for V_1V_{2S} structure, where V_1 is a semi vowel. In the case, when unstressed vowel of the diphthong is the last phoneme of the word and its visual appearance is very important for speech understanding, expressiveness coefficient for viseme of this semivowel is equal to 2/3 of vowels maximum expressiveness.

Finally two white rules in Fig. 3.11. describe the situation, when stressed diphthong is at the last syllable of word, which ends in consonant. We propose to apply linear interpolation between expressivity coefficients of these phonemes.

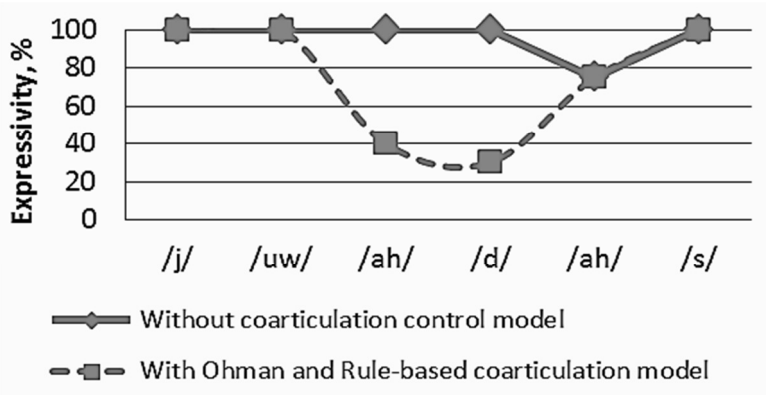


Fig. 3.11. Expressiveness parameter before and after application of our proposed model

To estimate quality of our proposed coarticulation control model, we've compared visemes expressivity before and after application of our model. Results of this experiment are shown in Fig. 3.11. They confirm that expressiveness parameters of visemes defined by proposed coarticulation model are much more reliable to coarticulation of Lithuanian word *"**juo**das"*, which includes pure Lithuanian diphthong.

3.2.3. Concluding remarks on coarticulation modelling for pure Lithuanian diphthongs

Since all humans are experts in lip reading and detects even the slightest errors during speech animation, expressive speech with integrated coarticulation rules is crucial part of any speech animation system. Ever since pure Lithuanian diphthongs regularly appear in Lithuanian words, their accurate visualization considerably improves the perspicuity of animated speech.

Eight rules for pure Lithuanian diphthong animation were defined in this paper. The proposed coarticulation model was applied in rule-based Lithuanian speech animation system. Comparison of visemes expressivity before and after integration of the proposed rules proved that intelligibility of animated Lithuanian words with pure Lithuanian diphthongs noticeably increased.

3.3. Practical evaluation of the proposed framework

The quality of Lithuanian „Talking head" is estimated in terms of intelligibility and naturalness. Since Lithuanian visual speech synthesizer is based on translingual adaptation, all the experiments are constructed to evaluate Lithuanian speech animation produced using English speech animation system „iFACE" and Lithuanian „Talking head" „LIT". Evaluation techniques similar to those used in Visual Speech Synthesis Challenge LIPS2008 (Theobald et al., 2008) were applied to evaluate Lithuanian speech animation framework.

3.3.1. Environment of the experiment

We pursued that experiment conditions would be as natural as possible. 23 persons with normal hearing and vision were individually asked to conduct 2 internet surveys, specially prepared for this experiment. All of them were Lithuanians of 17–30 years old. Characteristics like computer capacity, screen resolution or internet speed was not controlled, since usually people run the same application on different computers and expect equally positive results. However, all the participants were asked to wear headphones to perceive acoustic information.

Physical abilities like speechreading influence the alternation of the results as well as the experimental environment and its technical characteristics. Every hearing person has some capacity for speechreading, since they are better at understanding speech when they can see the talker, but they are worse when the talker's mouth movements do not correspond to what is heard. However, not all talkers are equal at speechreading. There is no compelling evidence for *systematic* talker differences based on age, gender, skin colour or perceived ethnicity (Bench et al., 1995). The best indicator of a talker's speech readability is *familiarity* with the talker, their speech style, and accent.

Moreover, speechreaders rest their gaze on the talker's face in the eye region, not on the talker's lips, whether they are viewing a live or recorded speech (Lansing et al., 2003). They can glance to the mouth region for specific information that might be predicted from the interpretation up to that point. It means that a lot of the information used in speechreading is not dependent on central vision processes, but includes peripheral vision processing too. This dynamic natural attention process cannot be simulated by zooming a camera plane of vision to bring it 'closer', because it simply interrupts human interpretation.

Campbell et al. (2010) defined the main factors that influences the speechreading ability, but cannot be simply estimated:

a) Lipreading *experience*. Historically, people who worked in noisy environments (machine weavers for example), became skilled lipreaders.

b) Good *language knowledge*, especially vocabulary skills. People are better at speechreading when they are familiar with the language being spoken.

c) Normal *vision*, with acute sensitivity to visual movement.

d) Good verbal *short term memory*. This contributes especially to sentence and discourse comprehension.

e) *Familiarity* with the talker, and the talker's accent and speech style.

f) *Risk-taking* personality traits (willingness to 'have a go', guessing).

g) *Age:* younger speechreaders (20–50 years) are better than older (>65 years), generally.

h) Gender, schooling, sign-language experience, scientific training have not been shown to affect speechreading skill reliably. People with *higher IQ* scores tend to be better speechreaders on some tests. Other personality traits (e.g. tenacity, self-confidence) and cognitive traits (e.g. executive function, attention switching, visual memory) are underexplored.

Ever since participants of our experiments had different suits of these qualities, both intelligibility and naturalness estimations noticeably varied between different participants.

3.3.2. Intelligibility testing of animated speech

Visual information is especially helpful in environments with degraded audio (Fagel et al., 2008). Besides, the perspicuity of clean audio is close to optimum either with or without video and visual-only lip-reading on semantically unpredictable sentences is almost impossible, even for lip-readers (Theobald, 2006). Therefore, for the experiments of the intelligibility, acoustic speech waveforms were firsly degraded with noise and later recombined with the videos for the experiments. The volume of the acoustic speech waveforms was firstly reduced by -10dB and mixed with Brown noise with maximum signal-to-noise ratio (SNR) of -30 dB. The amplitude of the noise level was chosen within a range in which the words were barely audible.

The same 3 groups of recorded Lithuanian words (in total 9 words) were presented in 3 distinct conditions:

a) degraded synthetic audio speech alone;

b) video that was generated using visemes, speech visualization engine and coarticulation rules originally implemented into „iFACE" system, specified to animate English language;

c) video that is the output of Lithuanian „Talking head" „LIT" (Fig. 2.8).

In each trial, participants were asked to transcribe orthographically what they believed they heard. In order to evaluate observer's hearing strength in noisy environments, viewers didn't have any list of probably pronounced words in audio standalone conditions.

For the other 2 conditions, participants had a list of possibly pronounced words and were requested to choose, which words were presented in video of animated speech. After each trial the subject was requested to select the most probable words. They didn't know how many words were pronounced in every trial. The words in each list were phonetically similar to the animated ones. For example, three Lithuanian words „*asilas, kapas, badas*" had to be identified from the larger list „*asilas, katinas, kapas, badas, karas, katilas, batas, padas, baras, darbas, vasara, kepalas*". Responses were scored as the number of words identified correctly.

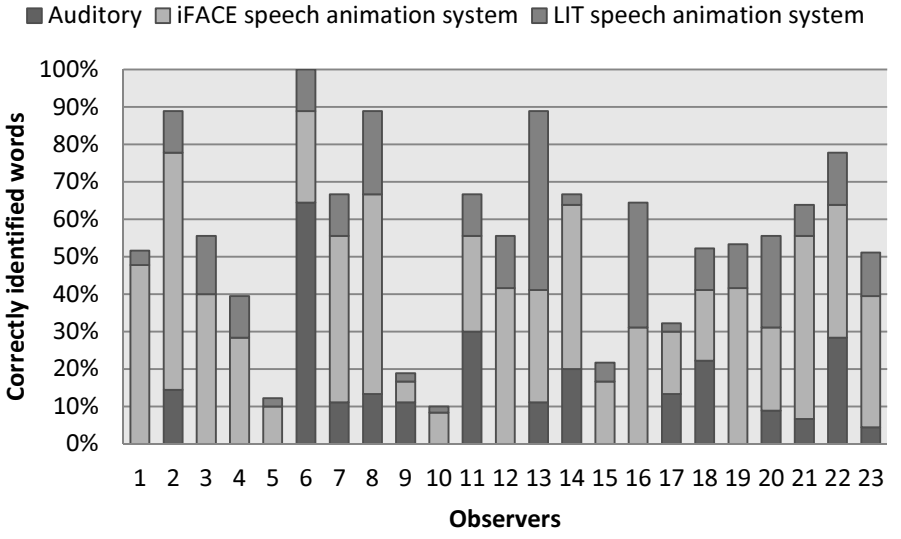


Fig. 3.12. The percentage of correct words for the different subjects in the three modalities (unimodal auditory, audiovisual speech using „iFACE" and „LIT" speech animation systems)

Final intelligibility results, that define the percentage of correctly recognized words in three distinct conditions described earlier, are presented in Fig. 3.12. Experiment results were calculated for each participant.
Visualization significantly improved the intelligibility of the animated Lithuanian speech for all 23 participants. Averages results of the estimates demonstrated in Fig. 3.12. are presented in Table 3.3.

Table 3.3. Average Lithuanian words recognition rate in in 3 distinct conditions

Evaluation environment	Words recognition rate, %
Auditory	**11**
iFACE system	**43**
LIT animation system	**56**

All the time the word recognition rate was higher for the audiovisual heads than for audio alone condition. Moreover, our proposed framework gave better Lithuanian speech intelligibility results comparing it with English speech animation system before its adaptation for Lithuanian language.

Under the unimodal auditory condition, the average of correctly recognized words was 11%. In audiovisual conditions Lithuanian words recognition rate increased by 32% using original implementation of „iFACE" system and by 45 % with our proposed framework. Despite subjectivity of participants and expected variation of technical experiment conditions, all 23 participants benefited from the proposed Lithuanian „Talking head".

3.3.3. Naturalness testing of animated speech

The naturalness tests were carried out after the intelligibility testing. Individual internet survey was prepared for this experiment. It involves playing synthesized video sequences lip-synched with auditory speech, where audio is not degraded at all. The clear acoustic speech is required for this type of evaluation, since it forms the basis for naturalness and believability judging. Besides, people are very sensitive to inaccuracies (both static and dynamic) in synthesized facial gestures: the overall lip-shape must be precise, the degree of articulation must be correct, and the auditory and the visual signals must be synchronised adequately. Frequently the viewer can detect that something in speech animation is erroneous, even if he cannot identify what exactly it is.

For this type of evaluation, the same audio file with recorded Lithuanian words was animated using English visual speech synthesizer „iFACE" and Lithuanian speech animation system „LIT". Outputs of these two different systems

were shown one after another. After the presentation of each video file, the observers were asked to rate the naturalness of the visual speech gestures along a five point Likert scale (Likert, 1932).

Distribution of the naturalness estimates collected during the experiment is summarized in Fig. 3.1.3. Values used to evaluate naturalness rating are:

5 – Speech animation is very natural

4 – Speech animation is natural

3 – Speech animation is neither natural nor unnatural

2 – Speech animation is unnatural

1 – Speech animation is very unnatural

All 23 participants judged the animation naturalness without knowing the recorded word sequences. They were asked to review the observed video three times before the final decision.

These 3 sets of recorded Lithuanian words were presented to all the participants:

a) 1 set. akt'i, k""ap's'i, kas""a, r""azdavo: (short /a/ in the first syllable, vowel at the end);

b) 2 set. ""a:Zuolas, prak""a:sto: (long /a:/ in long words with 3 syllables, specific Lithuanian phonemes included);

c) 3 set. ^a:s'ilas, k^a:pas, b^a:das (Accented Long /a:/ in the first syllable of the word, syllabic structure /as/ at the end of it).

Visual speech is significantly influenced by vowels and their position in the world. Moreover, the same vowel can be short and long. Therefore, 3 word sets with short and long /a/ in different places of long and short words were used to evaluate the overall naturalness of Lithuanian „Talking head". Averages of the results presented in Fig. 3.13 are revealed in Table 3.4.

The best naturalness results were achieved with the second set of words, since long words are easily recognizable both in auditory and audiovisual condition, however, it causes the lowest percentage of naturalness improvement (11,39%) comparing the original engine and our proposed adaptation of it. Improvement results on the other 2 sets with shorter words were significantly better (more than 18%). Speech animation was considered more as unnatural using visemes, speech visualization engine and coarticulation rules originally implemented into „iFACE" system, in the meantime Lithuanian „Talking head" was evaluated as neither natural nor unnatural.

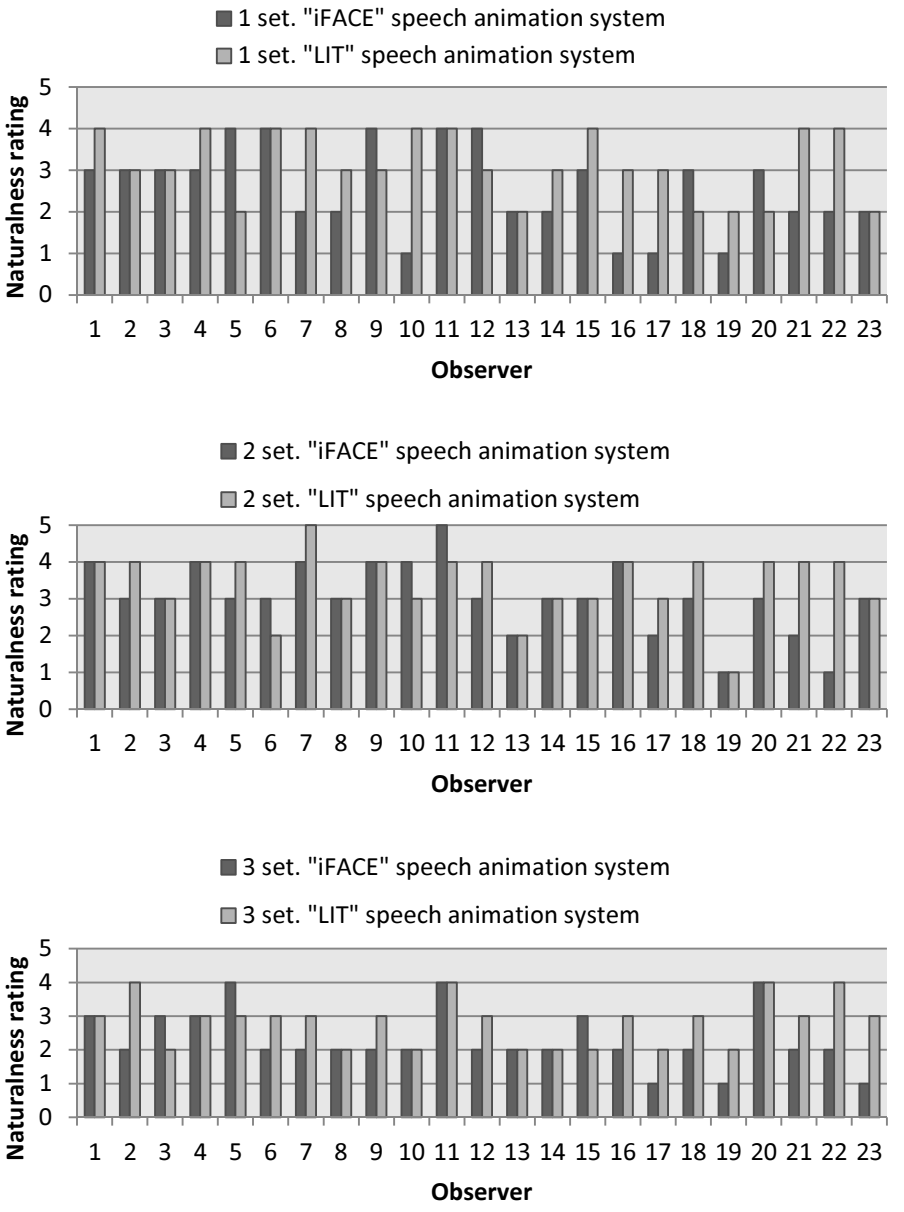


Fig. 3.13. Naturalness rating for the different subjects, when „iFACE" and „LIT" speech animation systems were applied to animate 3 sets of Lithuanian words

Table 3.4. An average naturalness result on videos generated with „iFACE" and „LIT" speech animation systems were applied to animate Lithuanian words. Likert skale, where very natural visual speech is scored as 5, was used for the estimation

Words set	iFACE system	LIT animation system	Improvement
1	2,56	3,13	18,15%
2	3,04	3,43	11,39%
3	2,30	2,83	18,46%
Average	**2,63**	**3,13**	**16,00%**

Viewers accented that that eye and head movements with face emotions would improve the perception of the animated speech. However, the concluding result shows that our framework (Fig. 2.8.) helps to improve the naturalness of visual Lithuanian speech by 16 %.

3.3.4. Concluding remarks on practical evaluation of the proposed framework

In this section practical experiments to evaluate the output of our proposed framework to animate recorded Lithuanian speech were presented. The quality of Lithuanian „Talking head" was estimated in terms of intelligibility and naturalness. For both types of the experiments quality of video samples generated with our proposed framework was compared with Lithuanian speech animation samples produced by speech animation system, which is based on English phonetics.

Our proposed framework improved the intelligibility of the lip animation for all 23 participants of the experiment. The intelligibility was evaluated in a word identification test. Lithuanian words recognition rate was estimated in 3 different conditions: audio alone as well as audiovisual using „iFACE" and „LIT" systems to animate 3 sets of Lithuanian words. Intelligibility evaluated in these surroundings was compared in between.

The naturalness of animated Lithuanian speech was estimated with 5 point Likert scale. These tests were designed to gauge the overall subjective opinion of how accurately animated speech corresponds to the acoustic speech. Improvement of naturalness results was calculated for 3 different sets of words. Even the naturalness results differ across individual participants; the concluding results show that our proposed framework helps to improve both intelligibility and naturalness of Lithuanian speech animation.

General Conclusions

In this thesis a novel framework to animate recorded Lithuanian speech was presented. Translingual speech animation technology was employed to create Lithuanian „Talking head" named „LIT". Theoretical evaluation and practical experiments allowed the following conclusions to be made:

1. State of art analysis revealed, that Lithuanian „Talking head" „LIT" should be viseme-driven speech animation system, which uses perspective parametric 3D head model to describe facial movements and to produce naturally looking animated speech.
2. The proposed new translingual speech animation technology allows division of viseme set of any language into constant (translingual) and variable parts. Based on the results of hierarchical clustering, 7 recommendations were offered to classify 22 internationally common consonants into 11 translingual viseme classes and to determine the constant part of visemes set. It is mutual for 11 analysed languages; therefore it can be used to animate other languages.
3. The proposed coarticulation control model allows simple integration of 8 Lithuanian phonetic rules to animate words with pure Lithuanian diphthongs. The original methodology that defines coarticulation modelling of the phonetic structures VVCV (vowel-vowel-consonant-vowel) was applied to attain speech naturalness. Coarticulation control

model specified for Lithuanian language helps to determine expressiveness parameter of specific visemes and to increase naturalness and intelligibility of animated Lithuanian speech.

4. The new speech animation framework, that integrates Lithuanian coarticulation control and translingual visemes classification models into chosen English speech animation system, was proposed to produce Lithuanian „Talking head" „LIT".
5. Lithuanian „Talking head" „LIT" was estimated in terms of intelligibility and naturalness during the experiments, where quality of video samples generated with „LIT" was compared with Lithuanian speech animation samples produced by English speech animation system „iFACE". The proposed Lithuanian speech animation technology improved speech animation naturalness by 13% and increased its intelligibility by 16%.

References

Abbasi, A. R., Ahmad, N. 2003. Urdu viseme identification, *Axbar-e Urdu* 19(6): 217–221.

Abry, C.; Boë, L. J. 1986. Laws' for lips, Speech *Communication* 5:97–104.

Ambrazas, V. 1997. Lithuanian Grammar, Vilnius: Baltos Lankos.

Arya, A.; DiPaola, S. 2004. Face As A Multimedia Object, in *Proc.5th International Workshop on Image Analysis for Multimedia Interactive Services, Lisbon, Portugal, April 21–23. 2004*. 21–23.

Arya, A.; DiPaola, S. 2007. Face Modeling and Animation Language for MPEG-4 XMT Framework, *IEEE Transactions on Multimedia* 9(6): 1137–1146.

Aschenberner, B.; Weiss, Ch. 2005. Phoneme-Viseme Mapping for German Video-Realistic Audio-Visual-Speech-Synthesis, *IKP – Working Paper NF 11, Institut für Kommunikationsforschung und Phonetik, Universität Bonn.*

Bailly, G.; Gibert, G.; Odisio, M. 2002. Evaluation of movement generation systems using the pointlight technique, *IEEE Workshop on Speech Synthesis, 2002.*

Balbonas, D.; Daunys, G. 2005. Classification of Phonemes using Sound and Visual Information, *Electronics and Electrical Engineering* 5(61): 74–77.

Balci, K.; Not, E.; Zancanaro, M.; Pianesi, F. 2007. 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*. 1013–1016.

Battista, S. 1999. MPEG-4: A Multimedia Standard for the Third Millennium, *IEEE Multimedia* 6(4).

Bench, J. 1995. Choosing talkers for the BKB/a speechreading test: a procedure with observations on talker age and gender, *British J. Audiology* 29:172–187.

Benoît, C.; Le Goff, B. 1998. Audio-visual speech synthesis from French text: Eight years of models, designs and evaluation at the ICP, *Speech Communication* 26:117–129.

Bescow, J. 1995. Rule-based Visual Speech Synthesis, in *Proc. of the 4th European Conference on peech Communication and Technology (Eurospeech'95), Madrid, Spain, 1995*. 299–302.

Beskow J. 1996. Talking Heads – Communication, Articulation and animation, in *Proceedings of Fonetik-96*. 53–56.

Bescow J. 2004. Trainable articulatory control models for visual speech synthesis, *Journal of Speech Technology*. 4(7):335–349.

Bevacqua, E.; Pelachaud, C. 2004. Expressive audio-visual speech, Journal *of Visualization and Computer Animation* 15(3–4): 297–304.

Black, A.; Bennett, C.; Blanchard, B.; Kominek, J.; Langner, B.; Prahallad, K.; Toth, A. 2007. CMU Blizzard 2007: A hybrid acoustic unit selection system from statistically predicted parameters, *Blizzard Challenge Workshop, Bonn, Germany, 2007*.

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

Bregler, C.; Covell, M.; Slaney, M. 1997. Video rewrite: driving visual speech with audio, in *Proc. of 24th conf. on Computer graphics and interactive techniques, ACM Press*. 353–360.

Breen, A.; Bowers, E.; Welsh, W. 1996. An Investigation into the Generation of Mouth Shapes for a Talking Head, in *Proceedings of ICSLP 96 (4), 1996*.

Brooke, M. N. 1996. Talking Heads and Speech Recognisers That can See: The computer Processing of Visual Speech Signals, in *Speechreading by Humans and Machines: Models, Systems and Applications*. 351–372.

Brooke, N. M.; Scott, S. D. 1998. Two and three-dimensional audio-visual speech synthesis, In *Proceedings of International Conference on Auditory-Visual Speech Processing (AVSP'98), Terrigal, Australia*. 213–218.

Bulterman, D. 2001. SMIL-2, *IEEE Multimedia*, 8(4).

Campbell, R; Dodd, B. 1987. *Hearing by Eye: The Psychology of Lip-Reading*. London, United Kingdom: Lawrence Erlbaum Associates.

Campbell, R.; Mohammed, T. 2010. Speechreading for information gathering: a survey of scientific sources. Deafness Cognition and Language (DCAL) Research Centre, Division of Psychology and Language Sciences, University College London.

Cao, Y.; Faloutsos, P.; Pighin, F. 2003. Unsupervised learning for speech motion editing, in *Proc. of ACM SIGGRAPH/Eurographics Symposium on Computer Animation, 2003.*

Cao, Y.; Faloutsos, P.; Kohler, E., Pighin, F. 2004. Real-time speech motion synthesis from recorded motions, in *Proc. of Symposium on Computer Animation, 2004.* 345–353.

Cassell, J. 2001. BEAT: the Behavior Expression Animation Toolkit, in *Proc ACM SIGGRAPH*.

Chen, T. 2001. Audiovisual speech processing. Lip reading and lip synchronization. *IEEE Signal Processing Magazine* 18:9–21.

Chuang, E.; Bregler, C. 2005. Moodswings: Expressive speech animation, *ACM Trans.on Graph.* 24(2).

Cohen, M. M.; Massaro, D. W. 1993. Modelling Coarticulation in Synthetic Visual Speech, *Models and Techniques in Computer Animation, Springer Verlag, Tokyo*. 139–156.

Cohen, M.; Massaro, D.; Clark, R. 2002. Training a talking head. IEEE Fourth International Conference on Multimodal Interfaces.

Cosatto, E.; Graf, H. P. 2000. Photo-realistic talking-heads from image samples, *IEEE Transactions on Multimedia* 2(3):152–163.

Cosatto, E. 2002. Sample-based talking-head synthesis. Ph.D. Thesis, Swiss Federal Institute of Technology,

Cosatto, E.; Ostermann, J.; Graf, H. P.; Schroeter, J. 2003. Lifelike Talking Faces for Interactive Services, *in Proc. of the IEEE, Special Issue on Human-Computer Multimodal Interface* 91(9): 1406–429.

Cosker, D.; Marshall, D.; Rosin, P.; Paddock, S.; Rushton, S. 2004. Towards perceptually realistic talking heads: models, metrics and McGurk, in *Proc. of Applied Perception in Graphics and Visualization, 2004.*

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

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

Damien, P.; Wakim, N.; Egea, M. 2009. Phoneme-viseme mapping for Modern, Classical Arabic language, in *Proc. of International Conference on Advances in Computational Tools for Engineering Applications, ACTEA '09, Lebanon, 2009.* 547−552.

DeCarlo, D.; Metaxas,D.; Stone, M. 1998. An anthropometric face model using variational technique, *in SIGGRAPH Proceedings, 1998.*

DeCarlo, B.; Bilvi, M.; Pelachaud, C. 2002. APML, a Markup Language for Believable Behaviour Generation, *First Intl Conf Autonomous Agents & Multi-Agent Systems, Workshop on Embodied Conversational Agents, Bologna, Italy, July 2002.*

DeMartino, J.M.; Magalhaes, L.P.; Violaro, F. 2006. Facial animation based on context-dependent visemes, *Computers and Graphics*, 30:971–980.

Deng, Z.; Neumann, U.; Lewis, J. P.; Kim, T. Y.; Bulut, M.; Narayanan, S. 2006. Expressive facial animation synthesis by learning speech co-articulations and expression spaces, *IEEE Trans. Vis. Graph.* 12(6):1523–1534.

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

Dipaola, S.; Arya, A. 2007. 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*. 129–136.

Dodd, B.; Campbell, R. 1987. *Hearing by Eye: The Psychology of Lipreading*, Lawrence Erlbaum, London, UK.

Elisei, F.; Odisio, M.; Bailly, G.; Badin, P. 2001. Creating and controlling video-realistic talking heads, in *Proceedings of the Workshop on Audiovisual Speech Processing (AVSP 2001). Scheelsminde, Denmark*. 90¬97.

Ekman, P.; Friesen, W. V. 1978. Facial Action Coding System, Consulting Psychologists Press.

Ezzat, T.; Poggio, T. 1998. Miketalk: A talking facial display based on morphing visemes, in *Proc. of Computer Animation Conference, 1998.*

Ezzat, T.; Geiger, G.; Poggio, T. 2002. Trainable video-realistic speech animation, in *Proc. of Computer Graphics and Interactive Techniques, ACM Press*. 388–398.

Fagel, S.; Bailly, G.; Elisei, F. 2007. Intelligibility of natural and 3D-cloned German speech, in *Proceedings of Auditory-Visual Speech Processing, 2007*.

Fagel, S. 2007. Auditory-visual integration in the perception of age in speech, *International Congress on Phonetic Sciences*. 725–728.

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

Filipovic, M. 2003b. Research of neural networks based isolated word recognition. *Lithuanian Mathematical Journal* 43:618–624 (in Lithuanian).

Fisher, G. G. 1968. Confusions amoung Visually Perceived Consonants. *Journal of Speech & Hearing Research* 15: 474–482.

Filipovic, M.; Lipeika, A. 2004. Development of HMM/Neural Network-Based Medium-Vocabulary Isolated-Word Lithuanian Speech Recognition System, *Informatica, Lith. Acad. Sci*. 15(4): 465–474.

Geiger, G.; Ezzat, T.; Poggio, T. 2003. Perceptual evaluation of video-realistic speech. CBCL Paper 224/AIMemo 2003003, Cambridge.

Goff, B. L.; Benoit, C. 1996. A text-to-audovisual-speech synthesizer for French, in *Proc. of the Int'l. Conf. on Spoken Language Processing (ICSLP).* 2163–2166.

Grammalidis, N.; Sarris, N.; Deligianni, F.; Strintzis, M. G. 2002. Three-Dimensional Facial Adaptation for MPEG-4 Talking Heads, in *EURASIP Journal on Applied Signal Processing, Special Issue on Signal Processing for 3D Imaging and Virtual Reality, 2002.* 10:1005–1020.

Hack, C.; Taylor, C. J. 2003. Modelling talking head behaviour, in *Proc. of British Machine Vision Conference, 2003.*

Hill, D. R.; Pearce, A.; Wyvill, B. 1988. Animating speech: an automated approach using speech synthesised by rules, *The Visual Computer* 3(4): 277–89.

Hirzalla, N.; Falchuk, B.; Karmouch, A. 1995. A Temporal Model for Interactive Multimedia Scenarios, *IEEE Multimedia*, 2(3):24–31.

Hoon, S. S. Chinese Speaking 3D Talking Head, Honours Year Project Report, National University of Singapore, 2003/2004.

Huang, F.J.; Chen, T. Real-time lip-synch face animation driven by human voice, in *Proc. of IEEE Multimedia Signal Processing Workshop, Los Angeles, California, 1998*. 352–357.

Hunt, A.; Black, A. 1996. Unit selection in a concatenative speech synthesis system using a large speech database, in *Proceedings ofICASSP'96*. 373–376.

Kasparaitis, P. 2008. Lithuanian Speech Recognition Using the English Recognizer, *Informatica* 19(4): 505¬516.

Kaukėnas, J.; Navickas, G.; Telksnys, L. 2006. Human-Computer Audiovisual Interface, *Information Technology And Control*, 35(2): 87–93.

Kelly, R.; Carson-Berndsen, J. 2006. Semi-Automatic Phonological Annotations of Speech by Grammatical Inference, in *Proc. of the Workshop on Annotation Science, 5th International Conference on Language Resources and Evaluation, Genoa, Italy, 2006*.1–8.

Kim, M.; Wood, S.; Cheok, L. T. 2000. Extensible MPEG-4 Textual Format (XMT), *ACM Conf Multimedia, 2000.*

Krnoul, Z.; Zelezn, M.; Mller, L.; Kanis, J. 2006. Training of coarticulation models using dominance functions and visual unit selection methods for audiovisual speech synthesis, in *INTERSPEECH-2006, 2006.*

Kshirsagar, S.; Thalmann, N. M. 2003. Visyllable based speech animation, *Computer Graphics Forum* 22(3).

Kuliešienė, D.; Grigonytė, G. 2005. The Potential of the Lithuanian Speech Corpus.

Kuratate, T.; Yehia, H.; Vatikiotis-Bateson, E. 1998. Kinematics-based synthesis of realistic talking faces, in *Proceedings of International Conference on Auditory-Visual Speech Processing (AVSP'98), Terrigal, Australia*. 185–190.

Lamel, L.F.; Kasel, R.H.; Seneff, S. 1987. Speech Database Development: Design and Analysis of the Acoustic-Phonetic Corpus, *in Proc.of the DARPA Speech Recognition Workshop, 1987*. 26–32.

Lansing, C. R.; McConkie, G. W. 2003. Word identification and eye fixation locations in visual and visual-plus-auditory presentations of spoken sentences, *Perception & Psychophysics* 65: 536–552.

Laurinčiukaitė, S.; Šilingas, D.; Skripkauskas, M.; Telksnys L. 2006. Lithuanian Continous Speech Corpus LRN 0.1: Design and Potential Applications, *Information Technology and Control* 35(4):431–440.

Laurinčiukaitė, S.; Filipovič, M.; Telksnys, L. 2009. Lithuanian Continuous Speech Corpus LRN 1: an Improvement, *Information Technology And Control* 38(3):203–207.

Lazalde, O. M.; Maddock, S.; Meredith, M. 2007. A Mexican-Spanish talking head, in *Proc. of the 3rd International Conference on Games Research and Development, CyberGames '07, Manchester Metropolitan University, UK, 2007.*17–24.

Le Goff, B.; Benoit, C. 1996. A Text-to-audiovisual speech Synthesizer for French, in *Proc.of ICSLP'96, Philadelphia, USA, 1996.*

Lipeika, A.; Lipeikiene, J.; Telksnys, L. 2002. Development of isolated word speech recognition system, *Informatica* 13(1):37–46.

Lofqvist, A. 1990. Speech as audible gestures, in *Speech Production and Speech Modelling*. 289–322.

MacDorman, K. F. 2006. Subjective ratings of robot video clips for human likeness, familiarity, and eeriness: An exploration of the uncanny valley. *ICCS/CogSci-2006 Long Symposium: Toward Social Mechanisms of Android Science,Vancouver, Canada.* 26–29.

Magno-Caldognetto, E.; Zmarich, C.; Cosi, P. 1998. Statistical definition of visual information for Italian vowels and consonants, in *International Conference on Auditory-Visual Speech Processing AVSP'98, Terrigal, Australia, 1998.*138–142.

Marriott, A.; Stallo, J. 2002. VHML: Uncertainties and Problems. A discussion, in *Proc First Int Conf Autonomous Agents & Multi-Agent Systems, Workshop on Embodied Conversational Agents, Bologna, Italy, July 2002.*

Massaro, D. W. 1998. Perceiving Talking Faces: From Speech Perception to a Behavioral Principle. *The MIT Press, Cambridge, Mass, USA*.

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

McGurk, H.; MacDonald, J. 1976. Hearing lips and seeing voices. *Nature* 264: 746–748.

Meyer, A. S.; Garchery, S.; Sannier, G.; Thalmann, N. M. 2003. Synthetic faces: Analysis and applications, *International Journal of Imaging Systems and Technology* 13(1):65–73.

Neto, J.; Cassaca, R.; Viveiros, M.; Mourão, M. 2006. Design of a Multimodal Input Interface for a Dialogue System, *Computational Processing of the Portuguese Language* 3960:170–179.

Ohman, S. 1967. Numerical model of coarticulation, *Journal of the Acoustical Society of America* 41: 310–320.

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

O´Neill, J.J. 1954. Contributions of the visual components of oral symbols to speech comprehension, *Journal of Speech and Hearing Disorder*, 19:429–439.

Ostermann, J.; Weissenfeld, A. 2004. Talking Faces – Technologies and Applications. *ICPR* 3: 826–833.

Ouni, S.; Massaro, D.; Cohen, M.; Young, K.; Jesse, A. 2003. Internationalization of a talking head, in *Proc. of 15th International Congress of Phonetic Sciences, Barcelona, Spain.*

Owens, E.; Blazek, B. 1985. Visemes observed by hearing-impaired and normal adult viewers, *Journal of Speech and Hearing Research* 28:381–93.

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

Parke, F. I. 1982. Parametrized models for facial animation, *IEEE Computer Graphics*, 2(9):61–68.

Parke, FI.; Waters, K. 1996.Computer Facial Animation. A K Peters, Ltd.

Pelachaud, C. E.; Magno-Caldognetto, Z; Cosi, P. 2001. Modelling an Italian Talking Head, In *Proc. Audio-Visual Speech Processing, 2001.* 72–77.

Pelachaud, C; Badler, NI; Steedman M. 1996. Generating facial expressions for speech, *Cognitive Science* 20(1): 1–46.

Picone, J. W. 1993. Signal modelling technique in speech recognition, in *Proc. of the IEEE*, 1993, 81(9): 1215–1247.

Pighin, F.; Hecker, J.; Lischinski, D.; Szeliski, R.; Salesin, D. H. Synthesizing Realistic Facial Expressions from Photographs, in *Proc.of the 25th Annual Conference on Computer Graphics, Orlando, FL, USA, 1998, July.* 75–80.

Raškinis, G.; D. Raškinienė. 2003. Parameter investigation and optimization for the Lithuanian HMM-based speech recognition system, in *Proc. of Information Technologies 2003, Technologija, Kaunas.* 41–48 (in Lithuanian).

Raškinis, G.; Raškiniene D. 2003b. Building medium-vocabulary isolated-word Lithuanian HMM speech recognition system, *Informatica* 14(1):75–84.

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

Recasens, D.; Pall`ares, M.; Fontdevila, J. 1997. A model of lingual coarticulation based on articulatory constraints, *Journal of the Acoustical Society of America* 102:544–561.

Reveret, L.; Bailly, G., Badin, P. 2000. Mother: a New Generation of Talking Heads Providing a Flexible Articulatory Control for Video-Realistic Speech Animation, in *Proceedings of the 6th International Conference on Spoken Language Processing (ICSLP'2000), Bejing, 2000*. 755–758.

Roach, P. 2004. British English: Received Pronunciation, *Journal of the International Phonetic Association* 34(2):239–245.

Rubin, P.; Baer, T.; Mermelstein, P. 1981. An articulatory synthesizer for perceptual research, *Journal of the Acoustical Society of America* 70 (2): 321–328.

Rudžionis, A.; Rudžionis, V. 1999. Phoneme recognition in fixed context using regularized discriminant analysis, in *Proc. of 6th European Conference on Speech Communication and Technology Eurospeech' 99.* 2745–2748.

Rudžionis, A.; Rudžionis, V. 2002. Lithuanian speech database LTDIGITS, in *Proc. of LREC 2002, Las Palmas, Spain, 2002*. 877–882.

Salleh, S. S.; Rahmat, R. W.; Mahmod, R.; Ahmad, F. 2010. 3D Lips Development and Measurement for Visual Speech Synthesis, *European Journal of Scientific Research* 35(2):159–172.

Sifakis, E.; Selle, A.; Mosher, A. R.; Fedkiw, R. 2006. Simulating speech with a physics-based facial muscle model, in *Proc. of Symposium on Computer Animation (SCA), 2006*.

Strom N. 1997. Phoneme probability estimation with Dynamic Sparsely Connected Artificial Neural Networks, *The Free Speech Journal* 1(5).

Šilingas, D.; Raškinis, G.; and Telksnys, L. 2003. Speech and Language Processing for Lithuanian: a Review, in *Proc. of Speech Processing Workshop, Magdeburg, Germany*, 2003. 57–64.

Šilingas, D.; Raškinis, G.; Telksnys, L. 2004. Review of Lithuanian Speech and Language Processing, in *Proc. of Human Language Technologies – the Baltic Perspective 2004, Riga, 2004.* 144–150.

Šilingas, D.; Laurinčiukaitė, S.; Telksnys, L. 2004. Towards Acoustic Modeling of Lithuanian Speech, in *Proceedings of SPECOM 2004, St. Petersburg: SPIIRAS, 2004.* 326-332.

Šilingas, D.; Laurinčiukaitė, S.; Telksnys, L. 2006. A Technique for Choosing Efficient Acoustic Modeling Units for Lithuanian Continuous Speech Recognition, in *Proc. of SPECOM 2006, St. Petersburg: SPIIRAS, 2006*. 61–66.

Tamulevicius, G.; Lipeika, A. 2004. Dynamic time warping based speech recognition system, in *Proceedings of The First Baltic Conference on Human Language Technologies: The Baltic Perspective, Riga.* 156–161.

Tekalp, A. M.; Ostermann, J. 2000. Face and 2-D mesh animation in MPEG-4. *Signal Processing: Image Communication* 15(4): 387–421.

Theobald, B.; Cawley, G.; Glauert, J.; Bangham, A. 2003. 2.5d visual speech synthesis using appearance models, in *Proc. ofBMVC 2003.* 1: 43–52.

Theobald, B.; Bangham, J. A.; Matthews, I.; Cawley, G. 2004. Near-videorealistic Synthetic Talking Faces: Implementation and Evaluation, *Speech Communication* 44:127–140.

Theobald, B.; Fagel, S.; Bailly, G.; Elisei, F. 2008. LIPS2008: Visual Speech Synthesis Challenge, in *Proceedings of Interspeech*.1875–1878.

Verma, A.; Subramaniam, V.; Rajput, N.; Neti, C. 2003. Animating Expressive Faces Across Languages, *IEEE Trans. on Multimedia* 6(6): 791–800.

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

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

Waters, K. 1987. A muscle model for animating three-dimensional facial expressions. *In proc. of Computer Graphics ACM SIGGRAPH'87.*

Waters, K.; Levergood, T. M. 1993. DECface: an automatic lip-synchronization algorithm for synthetic faces. Technical Report Series, DEC Cambridge Research Laboratory; September 1993.

Wu, Z.; Zhang, S.; Cai, L.; Meng, H. M. 2006. Real-time synthesis of Chinese visual speech and facial expressions using MPEG-4 FAP features in a three-dimensional avatar, in *INTERSPEECH'2006, 2006*.

Xie, L.; Liu, Z. Q. 2006. An Articulatory Appraoch to Video-Realistic Mouth Animation, in *Proc. of ICASSP, 2006.* 593–596.

Young, S. 1995. Large Vocabulary Continuous Speech Recognition: a Review, in *Proc. of Workshop on Speech Recognition, Snowbird, UT, USA, 1995*.

Young, S.; Kershaw, D.; Odell, J.; Ollason, D.; Valtchev V.; Woodland, P. 2000. The HTK Book. http://htk.eng.cam.ac.uk/docs/docs.shtml

Zailskas, V. 2011. Master Thesis: Lithuanian speech visemes visualization, Kaunas, gegužės 30d.

Zhang, Q.; Liu, Z.; Guo, B.; Shum, H. 2003. Geometry-driven photorealistic facial expression synthesis, in *Proc. of Symposium on Computer Animation.* 177–186.

List of Publications by the Author on the Topic of the Dissertation

Papers in the Reviewed Scientific Journals

Mažonavičiūtė, I.; Baušys, R. 2011. Translingual visemes mapping for Lithuanian speech animation, *Elektronika ir elektrotechnika* 5(111): 95–9. (ISI Web of Science).

Mažonavičiūtė, I.; Baušys, R. 2009. English talking head adaptation for Lithuanian speech animation, *Information technology and control* 3(38): 217–224. (ISI Web of Science).

Other papers

Mažonavičiūtė, I.; Baušys, R. 2010. Framework for Lithuanian speech animation, *18th European Signal Processing Conference (EUSIPCO-2010), Aalborg, Denmark, Rugpjūčio 23–27, 2010.* Aalborg : EUSIPCO. 781-785.

Ingrida MAŽONAVIČIŪTĖ

LIETUVIŲ KALBOS ANIMAVIMO TECHNOLOGIJA TAIKANT TRIMATĮ VEIDO MODELĮ

Daktaro disertacija

Technologijos mokslai,
informatikos inžinerija (07T)

LITHUANIAN SPEECH ANIMATION TECHNOLOGY FOR 3D FACIAL MODEL

Doctoral Dissertation

Technological Sciences,
Informatic Engineering (07T)

2010 00 00. 10,0 sp. l. Tiražas 20 egz.
Vilniaus Gedimino technikos universiteto
leidykla „Technika“,
Saulėtekio al. 11, 10223 Vilnius,
http://leidykla.vgtu.lt
Spausdino UAB „Ciklonas“
J. Jasinskio g. 15, 01111 Vilnius