THAI LIP-SYNC: การสร้างภาพเคลื่อนไหวริมฝีปากตามเสียงพูดภาษาไทย

นายทวีศักดิ์ ชื่นสายชล

วิทยานิพนธ์นี้เป็นส่วนหนึ่งของการศึกษาตามหลักสูตรปริญญาวิศวกรรมศาสตรมหาบัณฑิต
สาขาวิชาวิศวกรรมคอมพิวเตอร์ ภาควิชาวิศวกรรมคอมพิวเตอร์
คณะวิศวกรรมศาสตร์ จุฬาลงกรณ์มหาวิทยาลัย
ปีการศึกษา 2554

THAI LIP-SYNC: THAI SPEECH DRIVEN LIP ANIMATION

Mr. Thavesak Chuensaichol

A Thesis Submitted in Partial Fulfillment of the Requirements

for the Degree of Master of Engineering Program in Computer Engineering

Department of Computer Engineering

Faculty of Engineering

Chulalongkorn University

Academic Year 2011

Thesis Title	Thai Lip-Sync: Thai Speech Driven Lip Animation
By	Mr. Thavesak Chuensaichol
Field of Study	Computer Engineering
Thesis Advisor	Assistant Professor Pizzanu Kanongchaiyos, Ph.D.

Accepted by the Faculty of Engineering, Chulalongkorn University in Partial Fulfillment of the Requirements for the Master's Degree

…….. Dean of the Faculty of Engineering
(Associate Professor Boonsom Lerdhirunwong, Dr.Ing.)

THESIS COMMITTEE

…….. Chairman
(Assistant Professor Thanarat Chalidabhongse, Ph.D.)

……. Thesis Advisor
(Assistant Professor Pizzanu Kanongchaiyos, Ph.D.)

…….. Thesis Co-advisor
(Chai Wutiwiwatchai, Ph.D.)

…….. External Examiner
(Chutisant Kerdvibulvech, Ph.D.)

ทวีศักดิ์ ชื่นสายชล : THAI LIP-SYNC: การสร้างภาพเคลื่อนไหวริมฝีปากตามเสียงพูดภาษาไทย. (THAI LIP-SYNC: THAI SPEECH DRIVEN LIP ANIMATION)
อ. ที่ปรึกษาวิทยานิพนธ์หลัก : ผศ. ดร. พิษณุ คนองชัยยศ, 31 หน้า.

อุตสาหกรรมการสร้างแอนิเมชันมีได้รับความนิยมและมีการเติบโตไปอย่างมาก ซึ่งส่งผลให้ความต้องการที่จะเพิ่มประสิทธิภาพในการสร้างแอนิเมชันมีมากขึ้นโดยความต้องการที่จะลดระยะเวลาในการสร้างแอนิเมชันและลดภาระค่าใช้จ่ายในการสร้าง หนึ่งในขั้นตอนที่สำคัญที่สุดขั้นตอนหนึ่งคือการสร้างการเคลื่อนไหวริมฝีปากตามเสียงพูดให้กับตัวละครแอนิเมชัน โดยทั่วไปแล้วการสร้างภาพเคลื่อนไหวให้กับตัวละครแอนิเมชันจะกระทำในขั้นตอนสร้างการเคลื่อนไหวให้กับตัวละคร ในวิทยานิพนธ์นี้เราพิจารณาปัญหาของการสร้างการเคลื่อนไหวของริมฝีปากตามเสียงพูดของตัวละครเป็นหลัก จุดมุ่งหมายของวิทยานิพนธ์นี้คือการลดค่าใช้จ่ายและลดระยะเวลาในการสร้างการเคลื่อนไหวให้กับตัวละครแอนิเมชันที่พูดด้วยเสียงในภาษาไทย แนวคิดหลักของวิทยานิพนธ์คือการวิเคราะห์และระบุระยะเวลาของแต่ละหน่วยเสียงในการพูดของตัวละครและการเก็บข้อมูลการเคลื่อนไหวของริมฝีปากจากวิดีโอการพูดของมนุษย์ โดยขั้นตอนในการทำงานจะเริ่มต้นด้วยการแบ่งวิดีโอที่พูดด้วยมนุษย์ออกเป็นสองส่วนส่วนแรกคือส่วนที่มีการพูด โดยจะนำคำพูดรวมกับลำดับการพูดในแต่ละหน่วยเสียง เพื่อระบุระยะเวลาเริ่มต้นและสิ้นสุดของแต่ละหน่วยเสียงโดยใช้เทคนิคการระบุระยะเวลาหน่วยเสียง (Force Alignment) ซึ่งจะนำไปสร้างฐานข้อมูลการเคลื่อนไหวของริมฝีปากในแต่ละหน่วยเสียง (Visyllable Database) โดยการจับคู่การเคลื่อนไหวกับหน่วยเสียงนี้ทำได้โดยการนำข้อมูลเวลาเริ่มต้นของแต่ละหน่วยเสียงมาระบุภาพในวีดีโอซึ่งจะนำมาประกอบกับตำแหน่งที่สนใจบนใบหน้ามนุษย์และบันทึกตำแหน่งเป็นฐานข้อมูลการเคลื่อนไหวของริมฝีปาก จากนั้นจะสร้างส่วนหัวของตัวละครแอนิเมชันเป็นวิดีโอภาพเคลื่อนไหวได้โดยการนำข้อมูลระยะเวลาในการพูดของแต่ละหน่วยเสียงประกอบกับฐานข้อมูลในการเคลื่อนไหวของแต่ละหน่วยเสียง ซึ่งจากการทดลองและวิเคราะห์ผลการทดลองจะสามารถบ่งบอกถึงความถูกต้องของการเคลื่อนไหวได้ตรงกับคำพูดเมื่อเทียบเคียงกับภาพเคลื่อนไหวที่สร้างโดยศิลปิน

ภาควิชา...วิศวกรรมคอมพิวเตอร์... ลายมือชื่อนิสิต.....
สาขาวิชา...วิศวกรรมคอมพิวเตอร์... ลายมือชื่อ อ.ที่ปรึกษาวิทยานิพนธ์หลัก……..
ปีการศึกษา...2554.....

5370433221 : MAJOR COMPUTER ENGINEERING

KEYWORDS : LIP-SYNC / SPEECH DRIVEN ANIMATION / FORCE ALIGNMENT

THAVESAK CHUENSAICHOL : THAI LIP-SYNC: THAI SPEECH DRIVEN LIP ANIMATION. ADVISOR : ASST.PROF. PIZZANU KANONGCHAIYOS, Ph.D., 31 pp.

The animation industry is growing drastically. This results in increasing demand for better performance and reduction of process time and cost. One of the most important processes is lip synchronization. Generally the lip synchronization in character animation is done in the animation development process. In this research, we consider the problem of making lip movement for an animated talking character. We focus on to reducing the cost and workload in the animation development process, and apply this technique for use with Thai speech. The main idea is to extract and capture a viseme from the video of a human talking and the phonemic scripts inside this video. First, this approach starts with separating the human talking video into two parts that contains the speech and frame sequence, then uses speech combined with phonemic script to extract time-stamp of each phoneme by using force-alignment techniques; next, we create a visyllable database by mapping an end time of each selected phoneme to an image; then, we capture an interested position from the image to make a visyllable database; after that, we generate a talking head animation video by synchronizing a time-stamped of each phoneme to concatenated visemes. The output result of this research is the animation model that the animated talking character can move synchronously with the speech. The experiment reported, indicating good accuracy of the synchronized lip movement with the speech, compared to the artist-animated talking character.

Department : Computer Engineering Student's Signature .....

Field of Study : Computer Engineering Advisor's Signature .....

Academic Year : 2011

Acknowledgements

I would like to thank my thesis advisors: Asst. Prof. Dr. Pizzanu Kanongchaiyos and Dr. Chai Wutiwiwatchai for their advices and assistances, my thesis committee: Asst. Prof. Dr.Thanarat Chalidabhongse and Dr. Chutisant Kerdvibulvech for their comments and suggestions. We also would like to thank the Computer Graphics Laboratory at Chulalongkorn University (CUCG) for the helpful discussions and the Human Language technology (HLT) at National Electronics and Computer Technology Center (NECTEC) for the guidelines to speech processing technology. Finally, I deeply want to thank my parents for their love, understanding and invaluable supports throughout my study.

We acknowledge financial support from Thailand Graduate Institute of Science and Technology (TGIST) TG-44-09-53-063M and this work (AS585A) was partially supported by the Higher Education Research Promotion and National Research University Project of Thailand, Office of the Higher Education Commission. Special thanks go to Varakorn Ungvichian for his proofreading. The authors would like to thank Engineering Journal, Culture and Computing 2011 and ACM SIGGRAPH VRCAI'2011 reviewers for their comments and suggestions.

Contents

Page

List of Tables

Page

List of Figures

Page

CHAPTER 1

Introduction

Figure 1.1 Four snapshots from synthesized sequences. From the left to right : the snapshot pronounce /aa/, /uu/, /ii/ and /a/.

1.1 Introduction and Problem State

Recently, there has been widespread interest in character animation. Computer graphic technology has allowed us to create realistic characters that can follow human movement. Many technologies have matured for speech and video analysis, but the character expressions are still far from believable in most systems. Speech motions are generated by complex muscles around the face. We cannot justify the dynamics of these muscles nor their collaborative effects. Lip synchronization in character animation is generally done in animated films and games, consuming workload and cost during the animation development process. Generally, the interpretation of the lip motion data is done by a human observer who can specify its semantics. We develop a system for lip syncing the animated talking head characters from a visyllable database based on a speech-driven animation system.

The animator usually creates the lip synchronization by creating lip movements frame by frame, by observe the lip movements from human and adjust the movements as real as possible. However the quality and correctness of this process depends on the skill and performance of the animator. Some of the movement will not be correct as it should be. Furthermore, the movements will not follow the rule of how to pronounce in Thai.

We have considered the problem in this process and design to create the rigged structure and how to lip syncing the animated characters with Thai speech. The rigged structure is created by emulating the muscle-based around the face because we focus on the realism of the lip syncing and the correctness during the lip movement.

Generally, speech-driven speech animation uses the phonemes as the basic units of speech, and visemes as the basic units of animation. The process to create animation from the visyllable database and the phonemic time-stamped is separated into levels of processing in the following order: First, capture the viseme from interested marked position in images. Second, construct the visyllable database by using a sample-based model. Finally, the system produces a visual speech animation by recombining motion frames from the visyllable database by using time-stamped phonemic transcriptions. The frame per second and quality of the video is considered in this method because some of the Thai phonemes have a very short duration. If the duration of one frame is more than the phoneme duration, the video will have some appearance of conflicting lip sync.

As a result, our research is to create a system to make realistic talking characters from recombining the visyllable database with a time-stamp of each phoneme in speech: however, we do not intend to create a complete model of 3D animated characters. Our main purpose of this research is to create realistic talking-head characters without expressions and personalities of the characters.

1.2 Scope of Study

We have studied how to perform the lip synchronization for the animated character by using natural language processing and computer graphics technology. Our goal is to reduce the process of lip-syncing which is generally done by the animator and apply this technique to Thai speech. The proposed method is based on speech-driven animation techniques.

1.3 Expected Benefits

The development in this research will help to improve the quality of lip syncing in Thai speech by raising the quality and accuracy. Moreover, this research will decrease the workload and cost for animation industry in Thailand. In the educational

area, we can apply this research to make a tool for training the student and foreigner who want to study how to pronounce in Thai language.

1.4 Publications

T. Chuensaichol, P. Kanongchaiyos and C. Wutiwiwatchai, Thai Lip-sync: Mapping Lip Movement to Thai Speech in Engineering Journal, Vol. 3, No. 2, 2011 (in Thai).

T. Chuensaichol, P. Kanongchaiyos and C. Wutiwiwatchai, Thai Speech-driven Facial Animation, Accepted to be published in The second International Conference on Culture and Computing, (2011).

T. Chuensaichol, P. Kanongchaiyos and C. Wutiwiwatchai, Lip Synchronization from Thai Speech, Accepted to be published in The 10th International Conference on Virtual Reality Continuum and Its Applications in Industry, (2011).

1.5 Definition

1 **Phoneme**: the smallest segmental unit of sound in a language that is capable of conveying a distinction in meaning.

2 **Viseme**: a representation unit to classify speech sounds in the visual domain, based on the interpretation of the phoneme in the acoustic domain and describes the particular facial and lip movements that occur during the voicing of phonemes.

3 **Visyllable Database**: a set of recorded visemes that stores every viseme in the interest languages.

4 **Co-articulation**s: refers to the influence of a speech sound during the production of a preceding speech sound, the effects of which are seen during the production of sound.

5 **Speech Driven Animation**: a technique to create lip synchronization based on mapping the phonemes to visemes.

6 **Frame**: a still image or a snapshot of an animation. Basically, an animation is composed of several frames in a sequence.

7 **Force Alignment**: the speech processing techniques to extract time-stamps from speech given the speech track and phonetics transcription.

8 **Inverse Kinematics**: a technique to determining the parameters of a flexible joint to achieve a desired position.

1.6 Research Procedure

Our research is planned for a 10-month period starting in December 2010 and ending in August 2011. The process can be summarized into 7 stages as follows.

Task	Start	Duration (months)	12/10	1/11	2/11	3/11	4/11	5/11	6/11	7/11	8/11
Theory and literature reviews	Dec 10	4									
Application design	Feb 11	3									
Making a Visyllable Database	Apr 11	2									
Application Implementation	May 11	2									
Result evaluation	June 11	2									
Conclusion	July 11	1									
Thesis report	July 11	2									

Table 1.1 Research Procedure

CHAPTER 2

Literature Reviews

Speech driven animation has been implemented by commercial applications such as Autodesk Maya plugin, Blender plugin, Sitepal, Crazy Talk. These applications have become well known among the animation and game industry. In the research area, speech driven animation is still of interested and has been improved with many tools and techniques such as motion capture system, computer vision technology and human language processing. In this section, we review the commercial applications and related work in this research area.

2.1 Commercial Application

In the first section, we review the well known lip synchronization commercial applications. Each application has different strengths and supported formats but the approach in this application is still mostly related to the speech driven animation.

2.1.1 Autodesk Maya and Blender Plugins

The plugins in Blender and Autodesk Maya add a speech processing module to the main program. The plugins help the animator to sync the lip movement with the speech by showing the duration of each phoneme in speech. Unfortunately the plugins still do not support Thai speech.

2.1.2 Sitepal

Sitepal is known as internet based application that can generate an animated speaking character. The process in creating the speaking character begins with the user designing the animated character by picking a model from the Sitepal library or using the user's photo and then the user adds voice to the character. Finally, the application generates the speaking character with speech driven animation techniques based on text-to-speech system.

2.1.3 Crazy Talk

Crazy Talk is the lip syncing desktop application. The result of the Crazy Talk is a complete 2D animation environment. This application have an ability to editing the timeline and keyframe animation and can be extend to apply with 3D animation.

2.2 Related Works

In the second section, we review the research in the speech driven animation area. The review begins with the reviewing of the framework and the main process in speech driven animation. Most of the research is implemented from the [1] framework. The speech driven animation improves syncing quality by using the new speech processing techniques and enhances the realism of the animation by adding the expression of the animated character model, interpolation techniques and motion capture system.

2.2.1 DECface: An Automatic Lip-Synchronization Algorithm for Synthetic Faces [1]

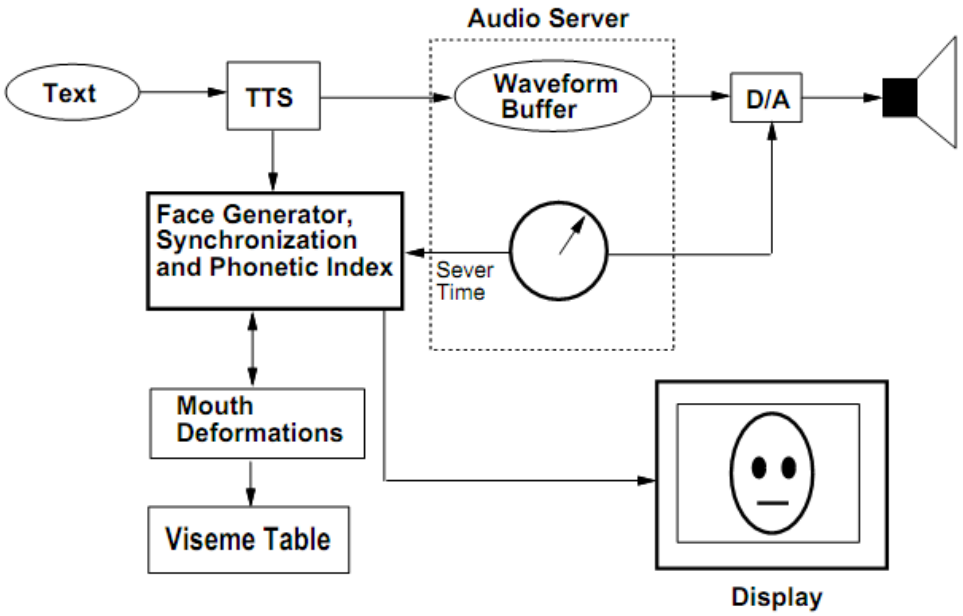

Figure 2.1 The synchronization model (image from: [1])

This research considers the problem of automatically synchronizing computer generated faces with synthetic speech. The goal of this research is to provide a novel form of face-to-face communication and the ability to create a talking personable synthetic character. This system is based on plain ASCII text input, a synthesized speech

segmentation module and synchronization in real-time to graphical display of an articulating mouth and face.

The speech input in this system is generated from automatic text-to-speech synthesis. Synthesizers in this system also accept input in the phonemes format. They separate the system into 3 parts: Formant-based rules programs, articulation-based rule programs and concatenation systems.

The algorithm in this research executes in the following sequence of operations:

1. Input ASCII text
2. Create phonetic transcription from the text
3. Generate synthesized speech samples from the text
4. Query the audio server and determine the current phoneme from the speech playback
5. Compute the current mouth shape from nodal trajectories
6. Play synthesized speech samples and synchronize the graphical display

The 2, 5 and 6 can be implemented and applied in our research. The phonetic transcription will guide the acoustic model to segment the speech. The step 5 and 6 are known as precisely techniques to create realistic lip syncing characters.

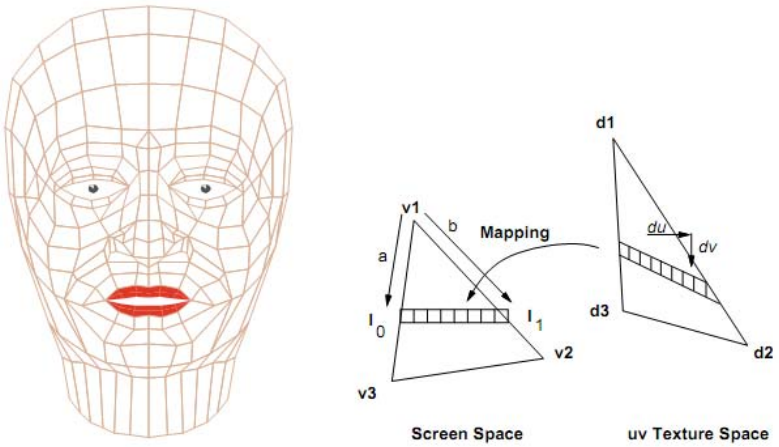

Figure 2.2 The left images show the polygonal representation of the face. the right image shows the texture mapping. (image from: [1])

2.2.2 Speech Driven Facial Animation [2]

This research describes the audiovisual system by learning the spatio-temporal relationship between speech acoustics and facial animation, including video and speech processing, pattern analysis, and MPEG-4 compliant facial animation for a given speaker.

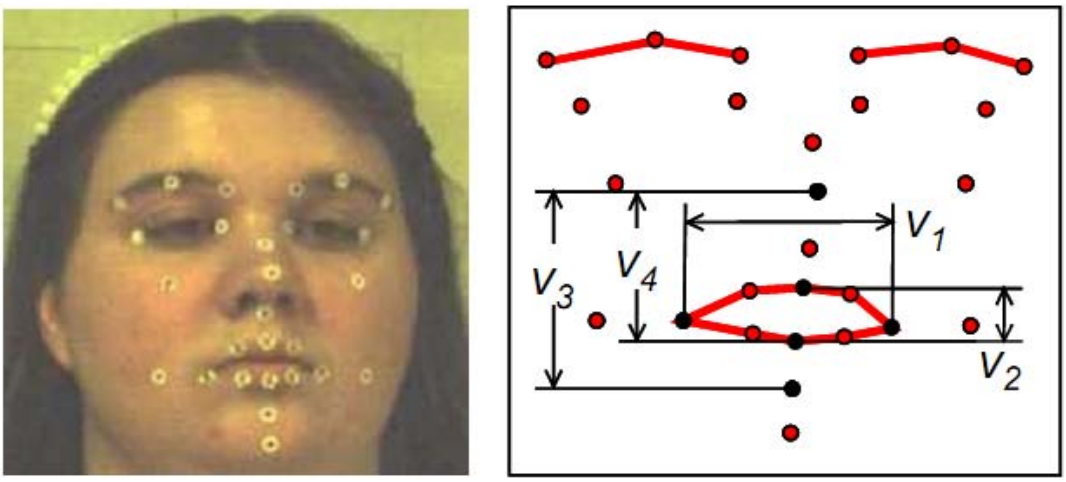

Figure 2.3 The neutral face of the subject with visual markers (image from: [2])

The video processing in this research uses the 3D tracking of the facial dynamics using a stereo camera pair and 27 markers placed

in the face. The position of each marker was independently tracked by finding the maximum cross-correlation in a local region of interest. The initial position of each marker was in the silence phoneme and neutral expression.

The audio-visual mapping techniques consider the acoustic features from past and future audio frames. The visyllable database was store in the database and be able to load if form of lookup table.

2.2.3 **Audio-Visual Synchrony for Detection of Monologues in Video Achieve** [3]

This paper presents the approach to detect the monologues in video shots. A monologue shot is defined as a shot containing a talking person in the video with the corresponding speech in audio. The goal of this research is to synchronize audio and face-based biometrics signals.

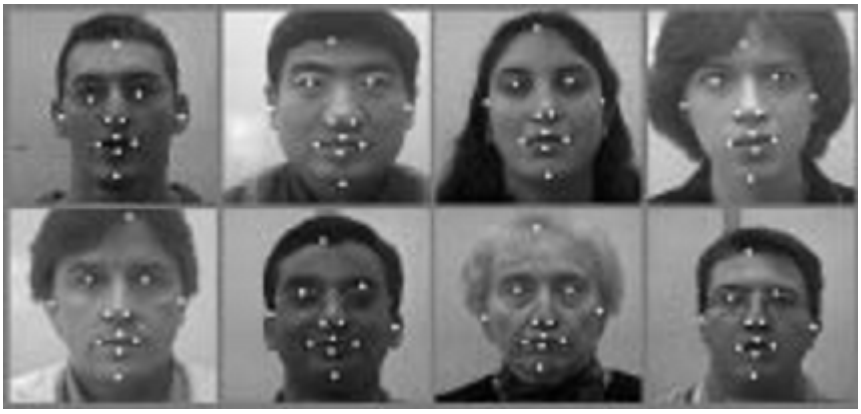

Figure 2.4 The ground-truth marked-up faces (image from: [3])

The face detector of their research uses likelihood ratio between two Gaussian Mixture models. They perform a 2D separable discrete cosine transform and retain the top 50 coefficients as the feature representation for these normalized faces.

2.2.4 Expressive Facial Animation Synthesis by Learning Speech Coarticulation and Expression Spaces [4]

This paper shows how to synthesize expressive facial animation by learning speech coarticulation models and expression spaces from recorded facial motion capture data. After users specify the input speech or texts and its expression type, the system automatically generates corresponding expressive facial animation.

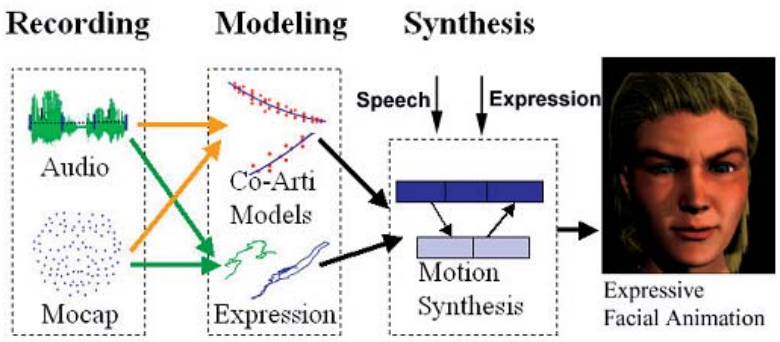

Figure 2.5 The main stage of the system (image from: **[4]**)

The synthesis system consists of two subsystems: neutral speech motion synthesis and dynamic expression synthesis.

Figure 2.6 The motion data acquisition (image from: [4])

The data acquisition and preprocessing were captured by a VICON motion capture system with a 120 Hz sampling rate.

The result of this paper is the synthesized expressive facial animation speech that learning from speech coarticulation model from real model capture. This system is efficient and reasonably effective but the limitation of this work is that it depends on the combinational of training data.

2.3 Summary of Literature Reviews

Most of the previous work on 3D characters lip synchronization based on speech-driven animation techniques implemented framework from [1]. However, this framework suffers from some limitations, because of input data that relied on text to speech system. Consequently, the improved framework from [2] was introduced. This framework proposes to generate animated talking characters by mapping phonemes in the speech track to the visyllable database, then recombining them to produce an animated talking video. The visyllable database for speech-driven animation can be captured from a motion capture system [3] [4] or 2D images [5]. According to the implementation of automatic speech recognition [6] in speech-driven animation, the segmentation can extract time-stamps from the speech audio track.

The advantage of this framework is good accuracy and flexibility to apply with other facial animation techniques, and with new languages.

In our previous research [7] we introduced the approach to extract time stamps by using force-alignment techniques with Thai speech. Some of our previous results have shown that most of the mismatched lip synchronization was mapped to final consonant phonemes of Thai speech. In this way, we enhance this previous approach to segmenting final consonants [8] in speech by adding speech and non-speech detector module. Additionally, we improve Thai speech-driven animation to support when use with multiple speaker force alignment techniques.

CHAPTER 3

Proposed Method and Implementation

This section shows the process of lip synchronization from our research. The result of the process produces an adaptively animated talking head character synchronized to the speech audio.

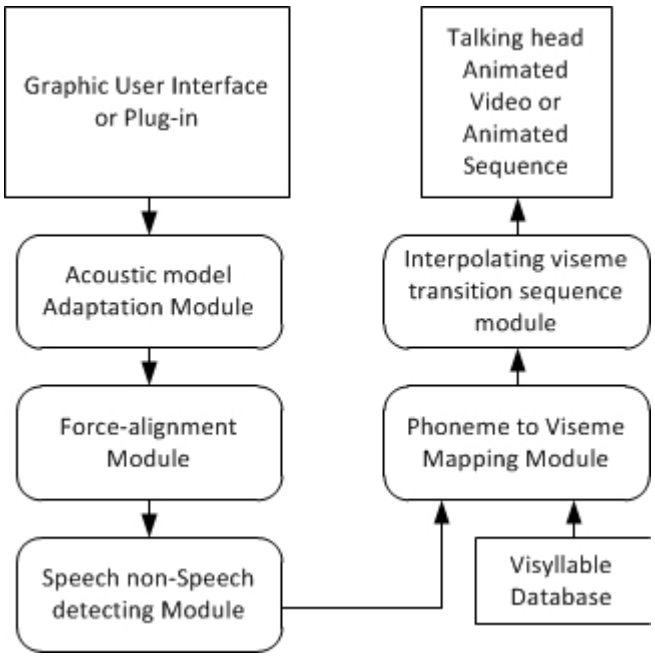

Figure 3.1 The Dataflow Diagram.

Figure 3.1 shows that our method has several advantages over the general method, by applying acoustic model adaptation and detecting final consonant phonemes. In addition, other existing methods do not fully support characteristics of the Thai language [8]

3.1 Speech-driven animation

Generally, speech-driven animation approaches synthesize new speech animations by concatenating visemic data. Their general processes are as follows: First, capture the viseme from 2D images or motion capture data. Second, construct the visyllable database by training a facial control unit with the machine-learning model or using a sample-based model. Finally, create a given visual speech animation by recombining motion frames from visyllable database by using time-stamped from phonetic transcriptions.

Co-articulation is an effect that can be observed during speech, in which facial movements correspond to one phonetic or visemic segment. In the process of articulating sentences, humans use preprocessing to create continuous speech. During these processes is the mixing of lip and jaw movements to compose phonemes and their transition. It is interesting that none of these models can explain co-articulation in different languages. In addition, there may not be a general model for co-articulation. Our approach is to interpolate and synchronize the lip movement in order of phonetic transcriptions. We would like to emphasize the advantages of our speech-driven animation system approach when compared to the complete known co-articulation effected speech-driven system. Additionally, this technique has a great benefit for Asian languages (Thai included) over the others lip syncing approach.

Speech-driven animation approaches typically generate realistic speech animation results, but it is complicated to select motion data to construct a balanced visyllable database for general animation characters.

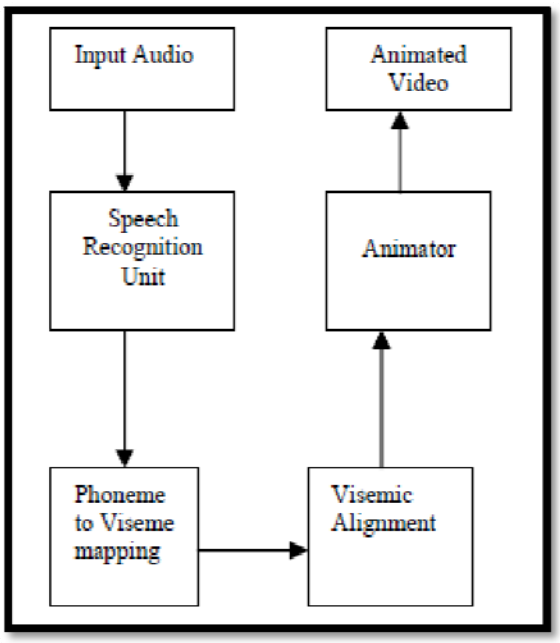

Figure 3.2 The Speech driven animation framework.

3.2 Force-alignment

HTK-based force-alignment [9] is one of the most reliable techniques to extract time-stamps from speech attached with the phonetic transcription. The different tonal phonemes in Thai speech do not affect the viseme [10].

Figure 3.3 shows the input, output and the process of the force alignment process. The speech track (MFCC format) is converted from speech track (wave format) with mfcc converter in HTK tools. The speech must start with the silence then follow with the speech and end with the silence in order. The dictionary contains all of the phonemes in the Thai language (sometime this file will include only the phonemes that appear in speech track). The phonetics transcription in Figure 3.3 means the phonetic dictionary in Thai language. It contains the phoneme sequences of the word. The HMM list (monophone) represents the form of the input and output of phonemes in force alignment process so in this research, we configure the same parameters as the dictionary file (In future work we can apply it to use with diphone and triphone). The training data is the speech track that we use to train the force-alignment acoustic model. The size and the amount of the data are varied to the length of speech

and number of speaker but the training data should at least cover all Thai phonemes. If the acoustic model has been trained before, we can skip the training session.

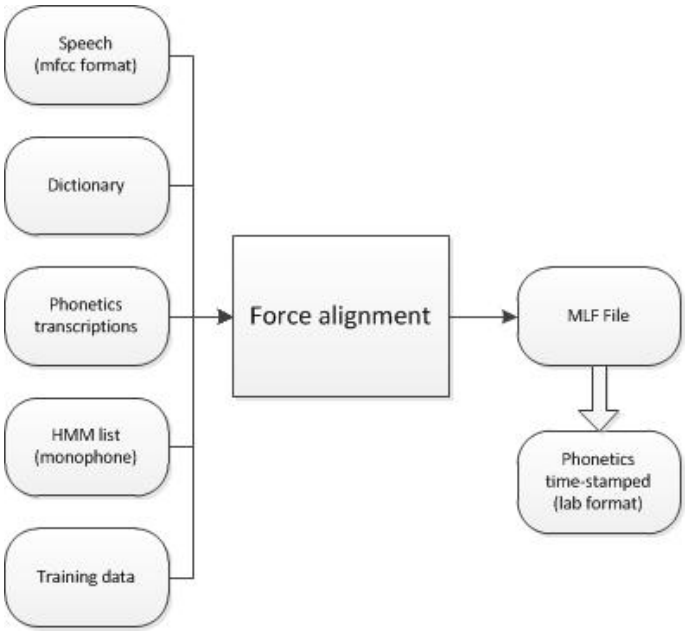

Figure 3.3 The force-alignment process.

The force-alignment process begins by converting the speech in the wave format to the MFCC format. The trained data is optional if the user wants to perform model adaptation before labeling the time-stamps. After the process is finished labeling, we divide the MLF file into individual speech (lab format) as shown in Figure 3.4.

```
0 17400000 sil
17400000 18400000 n
18400000 20500000 i
20500000 21800000 s
21800000 24000000 i
24000000 24300000 t^
24300000 25000000 c
25000000 26600000 uu
26600000 27700000 l
27700000 32200000 aa
32200000 55300000 sil
```

Figure 3.4 The result from the force-alignment process. The first two columns show the start and end time of the phoneme in the third column.

3.3 Personality of speaker

Regarding to [11] research is figures toward to contributing them preeminent behavior. In fact, personification means the act of attributing human characteristics to abstract ideas. The personification of the speaker affects the viseme capturing and force alignment process. The visyllable database uses to store the viseme. In the force-alignment process if the new speaker labels the time-stamps with other speakers adaptive acoustic models, the accuracy generally drops significantly. In viseme capturing, the speaker has some characteristics when talking that lead to the imbalanced visyllable database.

3.4 Acoustic model adaptation

In this research, we use the Thai language acoustic model from NECTEC [12]. The initial acoustic model is generated by statistically training based on Thai phoneme. Thai phonemes shown in Table 3.1 in CMU Thai phoneme from [13].

consonant	
single	combined
p	pr
t	phr
c	tr
k	kr
z	khr
ph	pl
th	phl
ch	thr
kh	kl
b	khl
d	kw
m	khw
n	transliterated
ng	br
l	bl
r	fr
f	fl
s	dr
h	
w	
j	

vow	
single	combined
a	ia
aa	iia
i	va
ii	vva
vow	ua
vv	uua
u	
uu	
e	
ee	
x	
xx	
o	
oo	
@	
@@	
q	
qq	

final consonant
p^
t^
k^
n^
m^
ng^
j^
w^
transliterated
f^
l^
s^
ch^

Table 3.1 List of Thai phoneme. (Image from: [13])

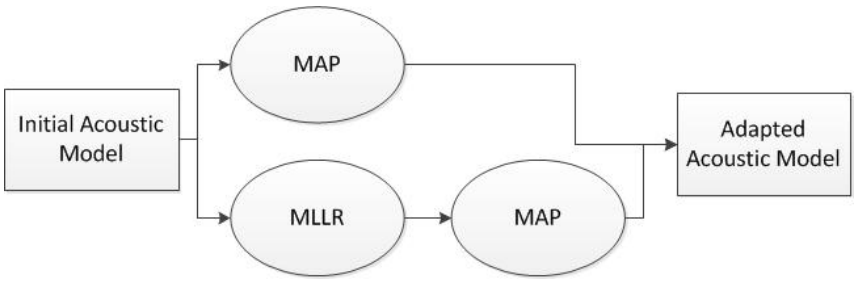

Figure 3.5 Acoustic model adaptation proceses.

We consider the appropriate technique to adapt our model. The well-known techniques that we choose to demonstrate are retraining technique, MAP, MLLR and MLLR+MAP. Our evaluation aims to find the effective adaptive techniques when the trained dataset is limited. From the second evaluation, we applied MAP (Maximum a posteriori estimation) [14] and MLLR (Maximum likelihood linear regression) model adaptation technique to our acoustic model.

According to the force alignment process, the model adaptation method can improve the accuracy of the acoustic model. In this way, the multiple speakers affect the size of trained data in model adaptation process. We found that the model adaptation method significantly improves force alignment process with multiple speakers. With regards to this problem, we develop a set of words to perform model adaptation method with the existing initial acoustic model. This trained data contains every phoneme in Thai speech and guarantees that each phoneme is used at least 15 times in the whole dataset.

3.5 Detecting the final consonant

As a result of the force alignment process, the time-stamps were extracted. A large error in time-labelling the time stamps of short final consonants are often observed due to the characteristics of the Thai language. Thus, we develop a second force-alignment module to detect speech and non-speech (/sil/ or /pau/ phoneme). Similarly to the previous force alignment module, the speech and non-speech force-alignment modules deliver a very reliable time-stamps extracting process. Furthermore, the extended module avoids adding time consumption, by using parallel

processing techniques (these two process are independent). The process in this method is executed in the following operations.

1. Input speech audio track
2. Attach speech and non-speech transcription
3. Extract speech and non-speech time-stamped
4. Determine the end point of final consonant in the speech
5. Merge the final consonant end point with previous results

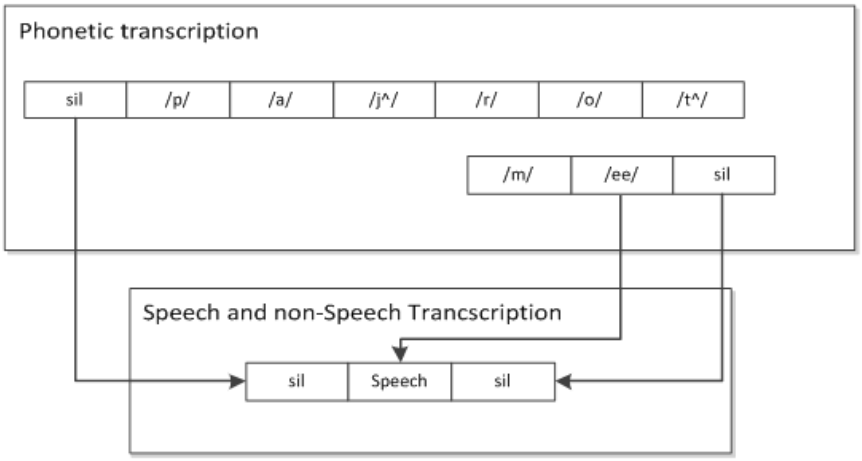

Figure 3.6 Mapping phonetic transcription to speech and non-speech (sil) transcription.

Figure 3.6 shows how to merge the time-stamp between phonetics and speech and non-speech force alignment process. The two silent phonemes were mapped together and /j^/ determined as final consonant of this speech.

Comparison of the merged time-stamps with the previous process indicates a better accuracy than the previous method using only force alignment system.

3.6 Visyllable Database

The characteristics of Thai language have some effect on the lip synchronization techniques. The number of phonemes is larger than English because

Thai has combined consonants, combined vowels and transliterated consonants (based on English) and transliterated vowels. However, the tone of Thai does not increase the size of visyllable database. In this way, we capture the viseme from the marked position in Figure 3.7. These marked positions will cover the movement of lip and jaw but the tongue and teeth is excluded because the capturing tool that we use does not support it, so we can only capture the external appearance of the human talking.

Figure 3.7 The marked position around the lip

We develop an approach to extract visemes that support a motion capture system or 2D images. The input data will be recorded and mapped into the corresponding phoneme. Generally, the viseme means the lip movement of the phoneme but in our design we change the representation of the data into the lip position to reduce the size of data and solve the co-articulation effect. To extract most of the consonant and final consonant viseme, we capture only the start position of the viseme. Unfortunately, some of the visemes cannot be captured accurately by this method because a single set of positions cannot describe the lip movement of such phonemes well enough. This group of phonemes contains combined vowels, combined consonants and the /r/ phoneme. Combined vowels and combined consonants can be easily captured by separating the capture position into 2 set of data but /r/ is the special case. The viseme of /r/ in Thai speech can observe as the tongue touch the upper platter repeatedly. To record and store the viseme data of /r/, we must record the 2 position of the start and end when tongue touch the platter and duplicate the moving of this effect repeatedly.

The visyllable database is collected by capturing from the marked positions. The audio and lip movements database was recorded from the user that read semantically neutral and expressing neutral state emotion at each reading. The facial markers were attached to the user face with the layout in the Figure 3.7. The capture camera sampling rate was set to 60 frames per second. The speech was recorded with a close talking BLUE Spark microphone at 48 KHz. After the data were collected, we use the force-alignment with the speech in the example video. After that, we capture the marked position from the frame matching the end time of each phoneme and store this data in the Euler's angle format. However, we transform the stored data into form of Quarternion angle.

The Quarternion angle is represented in the format shown in Equation

$$Quarternion\ angle = [a, b, c, d]$$

By combining the quaternion representations of the Euler rotations, we get the following equations. These equations transform the value of Eulur rotation into the Quarternion angle form. The values of this data from the equation represent the position of the bone in the rigged structure.

$$\mathrm{a} = \cos\frac{x}{2}\cos\frac{y}{2}\cos\frac{z}{2} + \sin\frac{x}{2}\sin\frac{y}{2}\sin\frac{z}{2}$$

$$\mathrm{b} = \sin\frac{x}{2}\cos\frac{y}{2}\cos\frac{z}{2} + \cos\frac{x}{2}\sin\frac{y}{2}\sin\frac{z}{2}$$

$$\mathrm{c} = \cos\frac{x}{2}\sin\frac{y}{2}\cos\frac{z}{2} + \sin\frac{x}{2}\cos\frac{y}{2}\sin\frac{z}{2}$$

$$\mathrm{d} = \cos\frac{x}{2}\cos\frac{y}{2}\sin\frac{z}{2} + \sin\frac{x}{2}\sin\frac{y}{2}\cos\frac{z}{2}$$

3.7 Rigging a Talking Head Model

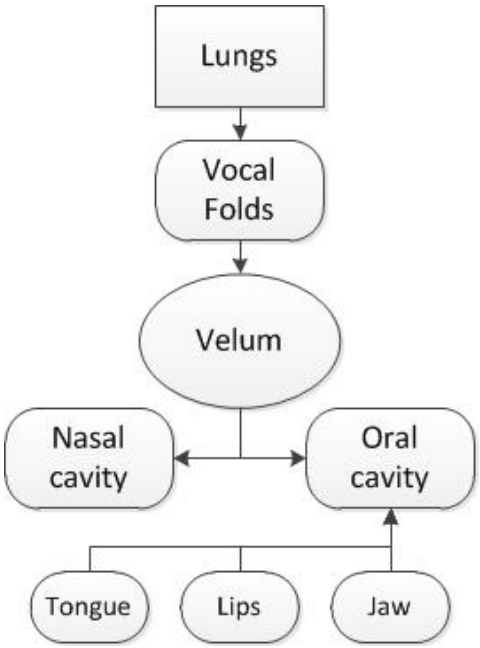

Figure 3.8 The schematic of air flow through the vocal tract

The properties of speech are still not completely understood. In the linguistics area concerns that the source of speech is created from the lungs, then pass through the vocal folds, the velum, lips, teeth and tongue as shown in Figure 3.8. The appearance of the vocal folds to the observer is from the lips, teeth, and tongue. The human face has an underlying skeletal structure. The main component of this structure is jaw. The surface around the skull is covered with muscles that can create movement. Generally, the muscles move the skin to the interested positions, and also produce a viseme. We rig our model based on muscles around the lip in human face as in Figure 3.9, excluding the tongue. This structure can perform a lip and jaw movement when the character is talking.

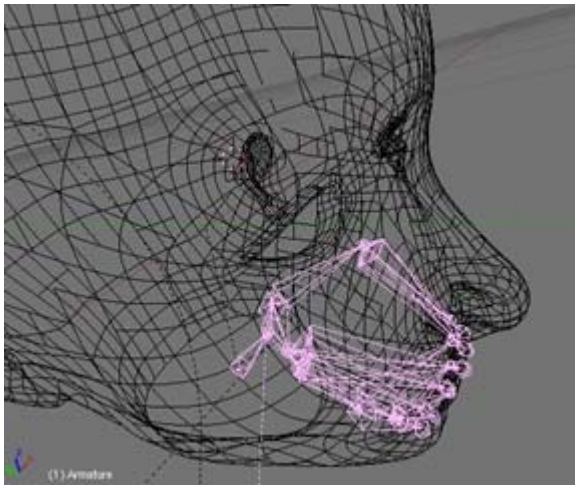

Figure 3.9 The rigged structure in talking head model

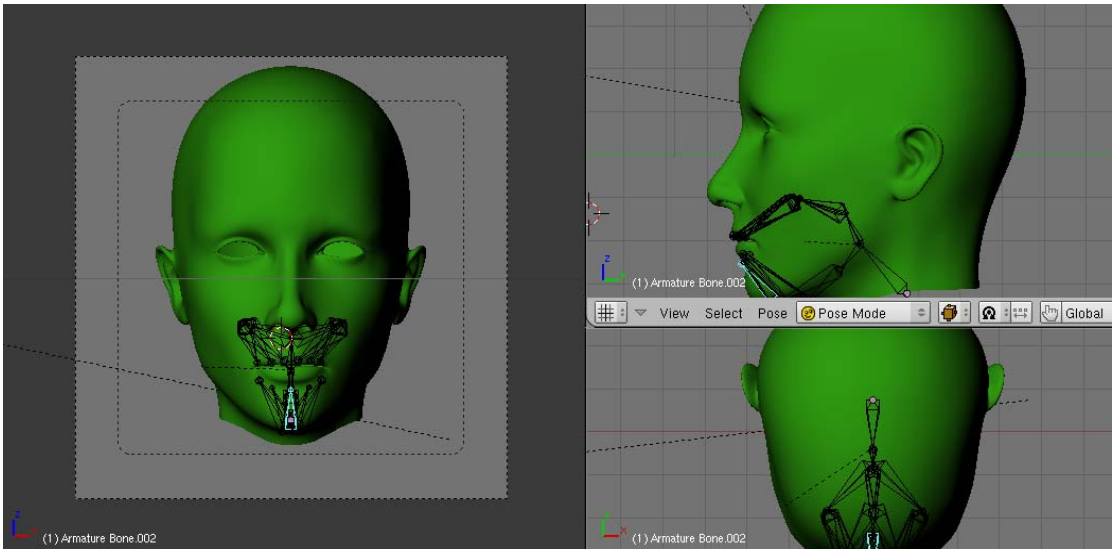

Figure 3.10 The rigged structure in three dimensions

3.8 Lip Motion Synthesis

To synthesize human-like lip motion, our technique aims for a realistic sequence of lip movement. We map lip movement from the visyllable database to the time-stamped in phonetic transcriptions. The end of the time-stamp in each phoneme is mapped to the key-frame in the visyllable database. In this process, we use inverse kinematic techniques to concatenate the co-articulation.

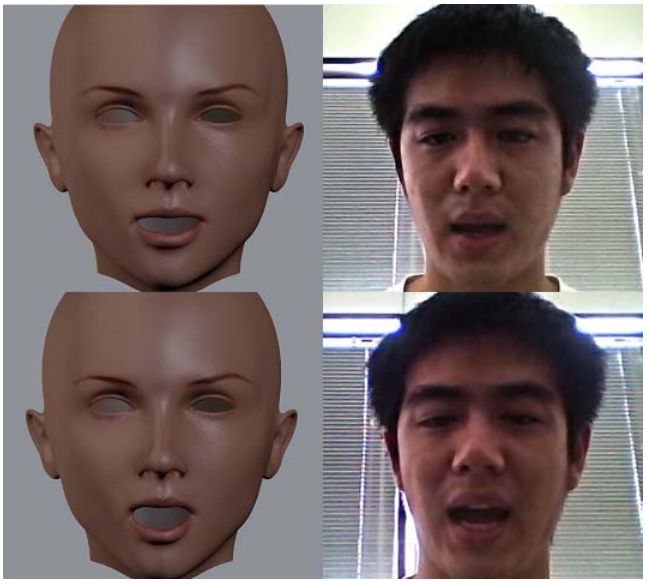

Figure 3.11 The two snapshots from synthesized sequences compared to the human: the first snapshots pronounce /ii/ and The second snapshot pronounce /aa/.

CHAPTER 4

Results and Discussion

To evaluate our system, we separate the experiment into three sections. In the first section, we measure the time labeling accuracy ability of the force-alignment processes. This experiment compares the various adapted acoustic model that contains the following techniques: the initial model, MLLR, MAP, MLLR+MAP and Retrain techniques by comparing the results of adapted acoustic model with manual labeling by human. First, we set the threshold value to 20, 40, 60, 80 and 100 milliseconds and then we count the number of phonemes that have an error of more than the threshold value. The threshold value in this experiment represents for a good quality of the video. The out-of-sync effect can be observed during viewing of the video if the error of lip sync is more than 80 milliseconds [15].

4.1 The accuracy of each model acoustic model adaptation method

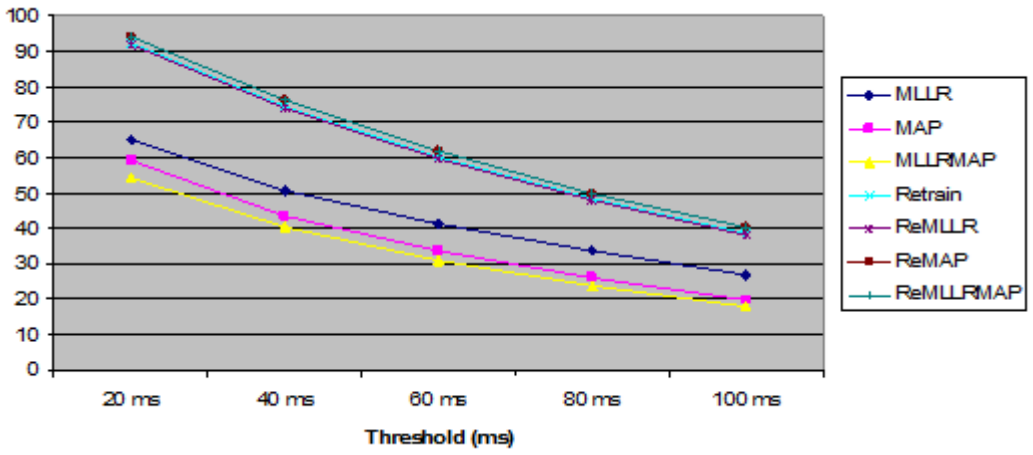

Figure 4.1 The comparison of the labeling accuracy of each adaptive acoustic model.

The benchmark in Figure 4.1 has shown that the MLLRMAP has the best labeling accuracy. Furthermore, we have tested the MLLR, MAP and MLLR+MAP in the small dataset. The results show that the MAP techniques significantly raise the accuracy when we train the acoustic model with 30 sentences (78 words), but MLLR

can raise the accuracy with any size of trained dataset. Traditionally, MLLR+MAP techniques slightly improve accuracy over both MLLR and MAP techniques without requiring a large trained dataset.

4.2 The accuracy of the labeling ability of acoustic model when apply speech and non-speech techniques.

In the second section, we measured the error /pau/ phoneme (same as /sil/ phoneme) with the adaptive acoustic model. Similar to the experiment in the first section, we set the threshold value to 20, 40, 60, 80 and 100 milliseconds.

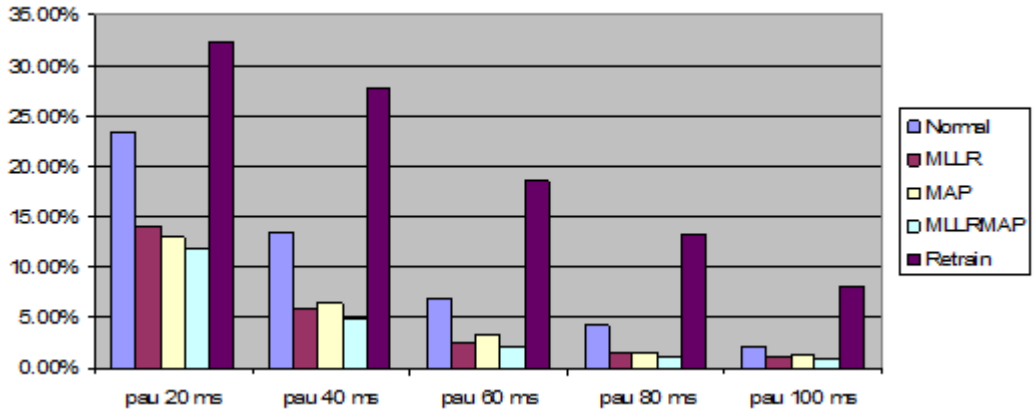

Figure 4.2 The comparison error in /pau/ phoneme of adaptive acoustic model.

Generally, the silence after speech is relative to the final consonant phoneme. Figure 4.2 shows the error of the silence is more than other phonemes. This experiment motivated us to develop the speech and non-speech detector module.

4.3 The quality of the animated video

In the final section, we measured the quality of video obtained from our speech-driven system by voting from 15 users. Every user watched the 10 synchronized videos and rated their quality.

The scale of evaluation is divided into:

5 = excellent, 4 = Very Good, 3 = Average, 2 = Fair, 1 = Poor.

Video	Score
Lab001.avi	4.67
Lab002.avi	4.53
Lab003.avi	5
Lab004.avi	4.13
Lab005.avi	5
Lab006.avi	4
Lab007.avi	4.33
Lab008.avi	4.26
Lab009.avi	4.13
Lab010.avi	4

Table 4.1 The voting results of the video quality.

CHAPTER 5

Conclusion and Future Work

In this section, we summarize the core ideas of our approach and the problems we found in this research and the future works.

5.1 Conclusion

According to the evaluation and results in Chapter 4.1 the MAP and MLLR+MAP adaptation method significantly improve acoustic model time-stamp extraction accuracy.

The acoustic model adaption can increase accuracy of the standard acoustic model in the system and be applied for use with the new user in the system without lowering the time-stamp extraction accuracy.

From the evaluation in Chapter 4.2 the speech and non-speech force alignment can improve the time-stamp extraction accuracy.

In this research, we have described how to develop speech-driven facial animation. The experimental data show that the model adaptation method and non-speech detector improve the lip synchronization quality. We have experimentally confirmed that the quality of the animated talking character is good.

5.2 Future Works

According from Chapter 4.3, the animated character can move with only lip and jaw, which is observed to be not very realistic. The teeth and tongue may be implemented to help improve the realism of animated character.

The expression of the character should be implemented because humans mostly produce speech that mix with the expression. This feature will also help improve the video quality and reduce the process in making animated characters.

References

[1] Waters, K., Levergood, T. M., and Laboratory, D. E. C. C. R., *DECface: An automatic lip-synchronization algorithm for synthetic faces*, Citeseer.

[2] Kakumanu, P., Gutierrez-Osuna, R., Esposito, A., Bryll, R., Goshtasby, A., and Garcia, O., Speech driven facial animation, Proceedings of the 2001 workshop on Perceptive user interfaces, ACM, pp. 1–5.

[3] Iyengar, G., Nock, H. J., and Neti, C., Audio-visual synchrony for detection of monologues in video archives, 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03)., Ieee, pp. V-772-5.

[4] Deng, Z. and others, Expressive Facial Animation Synthesis by Learning Speech Co-Articulation and Expression, Space, IEEE Transaction on Visualization and Computer Graphics, Citeseer, pp. 1-12.

[5] Matthews, I., Cootes, T. F., Bangham, J. a, Cox, S., and Harvey, R., Extraction of visual features for lipreading, Pattern Analysis and Machine Intelligence, IEEE Transactions on, 24 (2002): 198–213.

[6] Aggarwal, S. and Jindal, A., Comprehensive overview of various lip synchronization techniques, Biometrics and Security Technologies, 2008. ISBAST 2008. International Symposium on, IEEE, pp. 1–6.

[7] Chuensaichol, T., Kanongchaiyos, P., and Wutiwiwatchai, C., Thai Lip-Sync : Mapping Lip Movement to Thai Speech, Wisawakammasat Journal, (2011): 33-42.

[8] Chamnongthai, K., Final Consonant Segmentation for Thai syllable by Using Vowel Characteristics and Wavelet Packet Transform, ECTI Transactions on Computer and Information theory, 1 (2005): 50-62.

[9] Young, S., Gales, M., Liu, X. A., Woodland, P., Htk, T., and Version, H. T. K., *HTK Toolkit Book*, Cambridge University Engineering Department.

[10] Griffin, P. and Noot, H., FERSA: Lip-synchronous animation, Image Analysis Applications and Computer Graphics, (1995): 528–529.

[11] Joshi, P., Tien, W. C., Desbrun, M., and Pighin, F., Learning controls for blend shape based realistic facial animation, ACM SIGGRAPH 2005 Courses, ACM, p. 8–es.

[12] Thatphithakkul, N. and Kruatrachue, B., Denoise speech recognition based on wavelet transform using threshold estimation, Electrical Engineering Conference (EECON), Thailand (in Thai), pp. 2-5.

[13] Suebvisai, S., Charoenpornsawat, P., Black, A., Woszczyna, M., and Schultz, T., Thai automatic speech recognition, in Proc. ICASSP, IEEE, p. 857--860.

[14] Sharma, H. V. and Hasegawa-johnson, M., State-Transition Interpolation and MAP Adaptation for HMM-based Dysarthric Speech Recognition, Computational Linguistics, (2010): 72-79.

Biography

Thavesak Chuensaichol was born in Bangkok, Thailand, on 31 October, 1987. He began his schooling at the Yamsaard School, and then continued his education at Satriwitthaya 2 school, where his major's studies was physics and mathematics. In 2005, the year he gained his diploma, he entered Chulalongkorn University in Bangkok to be trained as an computer engineer. After four years of studied, he received his bachelor's degree and a scholarship for continuing master's degree in Computer Engineering, Chulalongkorn University.