Simulating Vocal Singing based on Speech Synthesis
Project Number 27

Attakorn	Benjadamrongrat	(Benz)	55070501056	hypertastic42010@gmail.com
Khantaphon	Chaiyo	(Tony)	55070503402	Khantaphon.c@mail.kmutt.ac.th
Thanyaporn	Saelim	(Kwan)	55070503419	kwan9888@hotmail.com

Advisor: Asst. Prof. Dr. Santitham Prom-on
Co-advisor: Asst. Prof. Dr. Suthathip Maneewongvatana

29 May 2016

____

I have read this report and approve its content.

Abstract

A text-to-speech (TTS) is the system that synthesizes artificial human voice. One possible use case of TTS systems is to create artificial singing voice. The critical challenges of singing synthesis concern with the synthesized sound naturalness and the pitch continuity. Thai TTS, however, was not widely known due to the lack of research and development in this area. Thai language is tonal language. Meaning of the spoken speech depends not only on the segmental parts of the spoken word but also on the tone. Thai language also has a complex orthography, where words cannot be simply separated by certain markers. So this project aims to develop Thai TTS for songs was made. A unit selection concatenative approach was implemented as a speech synthesis engine. Users need to specify Thai text, pitch specification in musical notes and preferred duration for each words. The system use dictionary-based word tokenization to split input text into words. Words, desired pitch and desired duration was used to select speech segment from pre-recorded database. For naturalness of synthesized voice, pitch contour of the synthesized speech are generated from recorded speech segments and desired pitch value. Speech segments are processed using TD-PSOLA to create voice with desired pitch and duration. Processed sound are concatenated together afterwards.

Acknowledgements

This project is successfully completed with get help from Asst. Prof. Dr. Santitham Prom-on and Asst. Prof. Dr. Suthathip Maneewongvatanathat spent their time for verify and check over our project. Giving advises and recommendation to improve the project which is very useful for develop this project. They also provide some of device to be our supply and create this project.

Thanks Computer Engineering Department, Faculty of Engineering, for workplace and tools for this project. We appreciate for all of teacher and professor that giving us knowledge that we needed to make this project completely and successfully and we also appreciate others more that giving help and knowledge , guide team ,and participate in making this successfully completed program .

Table of Contents

List of Figures

List of Tables

Chapter 1
Introduction

1.1 Problem Statement and Approach

Vocal Synthesis is the process to create sound of human voice using computer. Objective for vocal synthesis development varies on the purpose of usage. An announcement system on train stations and bank queue need clear, easy to hear voice. Such system use limited set of words and can be easily implemented. A generalized text-to-speech system, however, should be able to convert any text in a language into listenable voice with natural feeling. Having natural sounding voice similar to human voice would be a plus for text-to-speech system.

Recently songs that use synthesized vocal begin to popular. Songs that use synthesized vocal can overcome human speech limit in pitch and duration. Synthesized vocal also reduce the needs to hire professional singer. So composing music cost can be low enough to be personal hobby. Good vocal synthesizer for songs should be able to convert text in a language into human-like voice. Vocal synthesizer for songs also should be able to adjust pitch and duration, as well as some other voice characteristic such as tone and word stress.

Many available vocal synthesis available are in English and Japanese language. Availability of vocal synthesis system in Thai language is limited because of complex structure of Thai language. Thai language is tonal language, while English and Japanese is not. Thai vocal synthesis software for song is also not found in market. So if this type of software has been created, benefits to music composer community can be achieved. Vocal synthesis can also be reachable to public community who interested.

This project is belonged to research oriented and potential commercial product categories

1.2 Objectives

- Research the possibilities to create Thai language vocal synthesis system for singing
- Make vocal synthesis reachable by public who interested

1.3 Scope

The written software should have following capabilities

1. User input text to synthesize voice, which is based from pre-recorded database.
2. Using provided graphic user interface, synthesized voice can have its pitch and duration modified.
3. When user finished voice adjustment, the software allow user to export result as audio file.

1.4 Tasks and Schedule

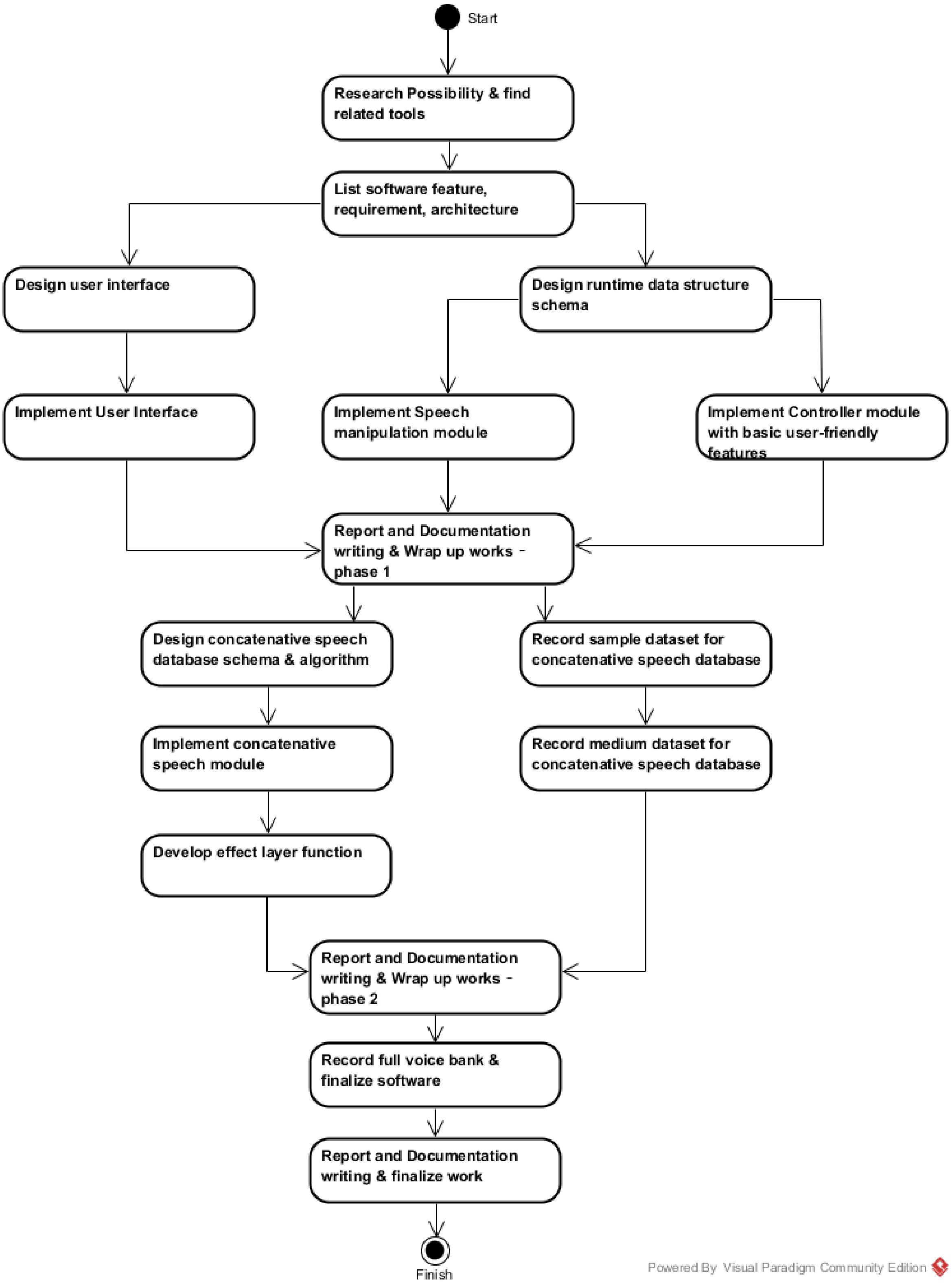

Start
Research Possibility & find related tools
List software feature, requirement, architecture
Design user interface
Design runtime data structure schema
Implement User Interface
Implement Speech manipulation module
Implement Controller module with basic user-friendly features
Report and Documentation writing & Wrap up works - phase 1
Design concatenative speech database schema & algorithm
Record sample dataset for concatenative speech database
Implement concatenative speech module
Record medium dataset for concatenative speech database
Develop effect layer function
Report and Documentation writing & Wrap up works - phase 2
Record full voice bank & finalize software
Report and Documentation writing & finalize work
Finish
Powered By Visual Paradigm Community Edition

Figure 1 Work Flow Diagram

Task Name	Time [Weekly]																																							
	August				September				October				November				December				January				February				March				April				May			
	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
Research Possibility & find related tools	■	■	■	■	■	■																																		
List software feature, requirement, architecture					■	■	■	■																																
Design user interface					■																																			
Design runtime data structure schema						■																																		
Implement User Interface							■	■	■																															
Implement Speech manipulation module										■	■																													
Implement Controller module with basic user-friendly features												■	■																											
Report and Documentation writing & Wrap up works – phase 1														■	■																									
Design concatenative speech database schema & algorithm																■	■																							
Record sample dataset for concatenative speech database																		■																						
Implement concatenative speech module																			■	■	■	■																		
Record medium dataset for concatenative speech database																							■	■																
Develop effect layer function																									■	■	■	■												
Report and Documentation writing & Wrap up works – phase 2																													■	■										
Record full voice bank & finalize software																															■	■	■	■	■	■				
Report and Documentation writing & finalize work																																					■	■		

Table 1 Project's Planning Gantt Chart

Chapter 2
Background, Theory and Related Research

2.1 Thai Language Writing and Pronunciation

Thai alphabet is syllabic, consisting of 44 basic consonants representing 21 distinct consonant sounds, each with an inherent vowel: [o] in medial position and [a] in final position. The [a] is usually found in words of Sanskrit, Pali or Khmer origin while the [o] is found native Thai words.

The 18 other vowels, diphthongs and triphthongs are indicated using diacritics which appear in front of, above, below of after the consonants they modify.

For some consonants there are multiple letters. Originally they represented separate sounds, but over the years the distinction between those sounds was lost and the letters were used instead to indicate tones.

Thai is a tonal language with 5 tones. The tone of a syllable is determined by a combination of the class of consonant, the type of syllable (open or closed), the tone marker and the length of the vowel. There are no spaces between words, instead spaces in a Thai text indicate the end of a clause or sentence and for the direction of writing in the Thai language is horizontal from left to right.

Consonants

Consonants are divided into three classes: middle, high and low, which help to determine the tone of a syllable. The sounds represented by some consonants change when they are used at the end of a syllable (indicated by the letters on the right of the slash below). Some consonants can only be used at the beginning of a syllable.

The following chart contains Thai consonants, their names and transcription

Symbol		Name	Symbol		Name
ก	k	kor kai	น	n	nor nuu
ข	kh/k	khor khai	บ	b/p	bor baimaai
ฃ	kh/k	khor khaut	ป	p	por plaa
ค	kh/k	khor khwai	ผ	ph	phor phueng
ฅ	kh/k	khor khon	ฝ	f	for faa
ฆ	kh/k	khor ra-khang	พ	ph/p	phor phaan
ง	ng	ngor nguu	ฟ	f/p	for fan
จ	j/t	jor jaan	ภ	ph/p	phor samphao
ฉ	ch	chor ching	ม	m	mor maa
ช	ch/t	chor chaang	ย	y	yor yak
ซ	s/t	sor soo	ร	r/n	ror runa
ฌ	ch	chor choe	ฤ	rue	ror rue(short)[1]
ญ	y/n	yor chin	ฤา	rue	ror rue (long)[1]
ฎ	d/t	dor chadaa	ล	l/n	lor/ling
ฏ	t	tor patak	ฦ	lue	lor lue (short)[1]
ฐ	th/t	thor santhaan	ฦา	lue	lor lue (long)[1]
ฑ	th/t	thor naangmonthoo	ว	w	wor waen
ฒ	th/t	thor phuuthao	ศ	s/t	sor saalaa
ณ	n	nor neen	ษ	s/t	sor reusii
ด	d/t	dor dek	ส	s/t	sor sela
ต	th/t	tor tao	ห	h	hor hiip
ถ	th/t	thor thung	ฬ	l/n	lor julaa
ท	th/t	thor thahaan	อ	[2]	or aang
ธ	th/t	thor thong	ฮ	h	hor nok-huuk

Table 2 Thai consonants, names and transcription

[1] Consonant-vowel combination characters, not members of any group.

[2] อ is a special case in that at the beginning of a word it is used as a silent initial for syllables that start with a vowel.

Vowels

The vowels each exist in long-short pairs: these are distinct phonemes forming unrelated words in Thai, but usually transliterated the same. The long-short pairs are as follows (a dash (–) indicates the position of the initial consonant after which the vowel is pronounced)

LONG			SHORT		
Thai		**Explanation**	**Thai**		**Explanation**
–า	a:	a in "father"	–ะ	a	u in "nut"
–ี	i:	ee in "see"	–ิ	i	y in "greedy"
–ู	u:	ue in "blue"	–ุ	u	oo in "look"
เ–	th e:	a in "lame"/t	เ–ะ	e	e in "set"
แ–	æ:	a in "ham"	แ–ะ	æ	a in "at"
–ื	ɨ:	u in French"dur" (long)	–ึ	ɨ	u in French "du" (short)
เ–อ	ə:	u in "burn" (long)	เ–อะ	ə	u in "burn" (short)
โ–	o:	ow in "bowl"	โ–ะ	o	oa in "boat"
–อ	ɔ:	aw in "raw"	เ–าะ	ɔ	o in "for"

Table 3 long and short pair Thai vowels

The basic vowels can be combined into diphthongs and triphthongs as follows

LONG			SHORT		
Thai		**Explanation**	**Thai**		**Explanation**
–าย	aːj	I in "I" (stressed)	ไ–, ใ–, ไ–ย	ɑj	I in "I"
–าว	aːw	ao in "Lao"	เ–า	aw	ow in "cow"
เ–ีย	iːa	ea in "ear" (long)	เ–ียะ	ia	ea in "ear"
–ิว	iw	ew in "new" (short)	–ัว	uːa	ewe in "newer"
–ัวะ	ua	ure in "pure" (short)	–ูย	uːj	ooee in "cooee!"
–ุย	uj	uey in "bluey"	เ–ว	eːw	a in "lame" + o in "poke"
เ–็ว	ew	e in "set" + o in "poke"	แ–ว	æːw	a in "ham" + o in "poke"
เ–ือ	ɨːa	u in French "dur" + a in "father"	เ–ย	əːj	u in "burn" + y in "yes"
–อย	ɔːj	oy in "boy" (long)	โ–ย	oːj	oe in "Chloe"

Table 4 Thai dipthongs vowels

LONG			SHORT		
Thai		**Explanation**	**Thai**		**Explanation**
เ–ียว	iow	ee + aow	–วย	uɛj	oo + I in "I"
เ–ือย	ɨɛj	u in French "dur" + I in "I"			

Table 5 Thai triphthongs vowels

2.2 Concatenative Speech Synthesis

Concatenative speech synthesis is a method to generate synthesized speech by combining recorded speech files from database. This method can provide many output without recording long speech but the quality of synthesized voice will be less natural than original speech itself.

2.3 Unit Selection

Unit selection is a method to query speech file from speech database by using the property of speech such as duration, pitch, content in the input sentence etc. to create the search target. The process is using speech database as state transition network and search by using each word to compare as send the most closest to the input. The closer of the result is, the easier to modify the word.

2.4 Time-Domain Pitch Synchronous Overlap-Add Algorithm (PSOLA)

TD-PSOLA (Time Domain Pitch Synchronous Overlap Add) is the method used to modify pitch and duration of speech sound but still remain the natural of speech. The process of TD-PSOLA are as follow

1. Marking the pitch in the speech. The input speech wave are marked periodically. The distance between marks are instantaneous period of pitch.
2. For each pitch points obtained, divided the sound using Hanning window with width equal to pitch's period. After this stage we'll get many segments of speech.
3. Overlap or extend the waves back into output wave. During this stage, we can adjust duration by duplicate or remove some segments. Adjusting pitch can also be done by shorten or lengthen distance between segments.

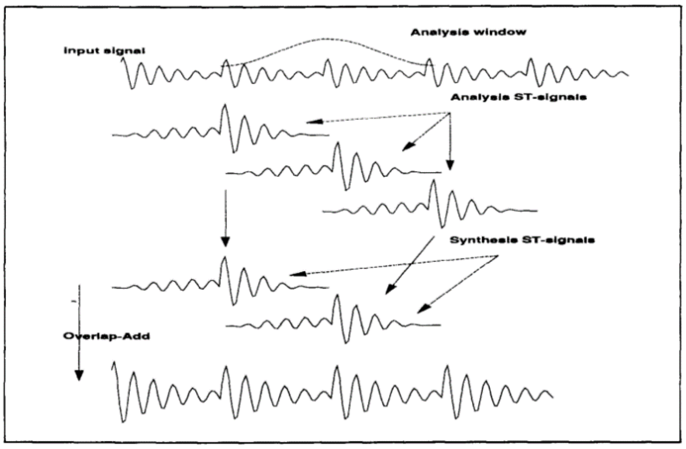


Figure 2 Duration reduction in PSOLA (H.Valbret,et al., 1992)

2.5 .NET Framework

.NET Framework is an application framework that developed by Microsoft. Both Windows application and web service can be developed by using .NET Framework.

The .NET Framework comprises of the basic language runtime and the .NET Framework class library. The normal language runtime provided hybrid implementation of compiler and interpreter. .NET Framework class provides interface to prewritten library. During application deployment, .NET Framework classes are packed as a redistributed package, which is specific to machine architecture and sharable with other application that developed in the same way.

2.6 Windows Presentation Foundation

WPF is stand for Windows Presentation Foundation, which is a Graphic User Interface framework that can be used with .NET framework. To create a GUI page, each component are arranged as hierarchy structure, which are saved in an XAML file. Each XAML file are coupled with a C# source code to provide event based interaction. So in the MVC paradigm, WPF work as VIEW part of the architecture.

2.7 Word Segmentation

Word segmentation is the process of divide string into component words. In Thai language, the problem is quite complex to solve because delimiter is not found in the text. Word segmentation approach currently available are dictionary based and machine learning based. Machine learning based word segmentation are under research, especially for Thai language.

Lexto is the word segmentation program that use dictionary based approach developed by NECTEC. However, one major problem of the dictionary approach is some words are substring of another words. Since this project aim to synthesize the voice, a smaller dictionary consists of only short words is needed for domain specific problem.

Chapter 3
Design and Methodology

3.1 Study of Difference between Normal Voice and Singing Voice

An experiment was conducted to find difference between normal speaking voice and singing voice. The experiment procedure are

1. Record speaking voice and singing voice on same sentence
2. Cut both speaking voice and singing voice into segments, each segment corresponds to one syllable
3. Using Praat software, find any interesting difference in pitch contour

3.2 Vocal Synthesis for Singing Software Requirements

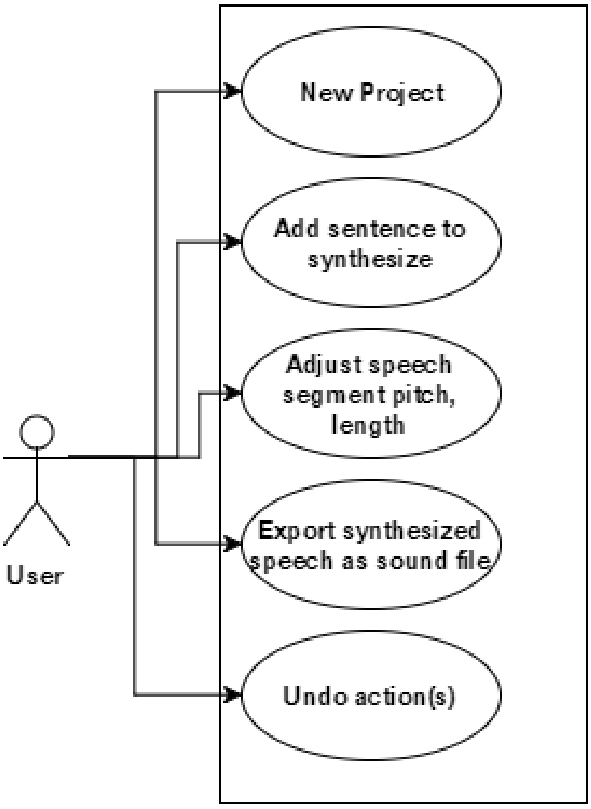

New Project
Add sentence to synthesize
Adjust speech segment pitch, length
Export synthesized speech as sound file
Undo action(s)
User

Figure 3 Software use cases diagram

User will be able to input sentence(s), which will be converted in to speech segment object show on program's user interface. User will be able to adjust speech segment properties such as pitch and length. User can undo their changes. User can save and load their work in progress and can be exported into sound file when project is completed.

3.3 Software Architecture

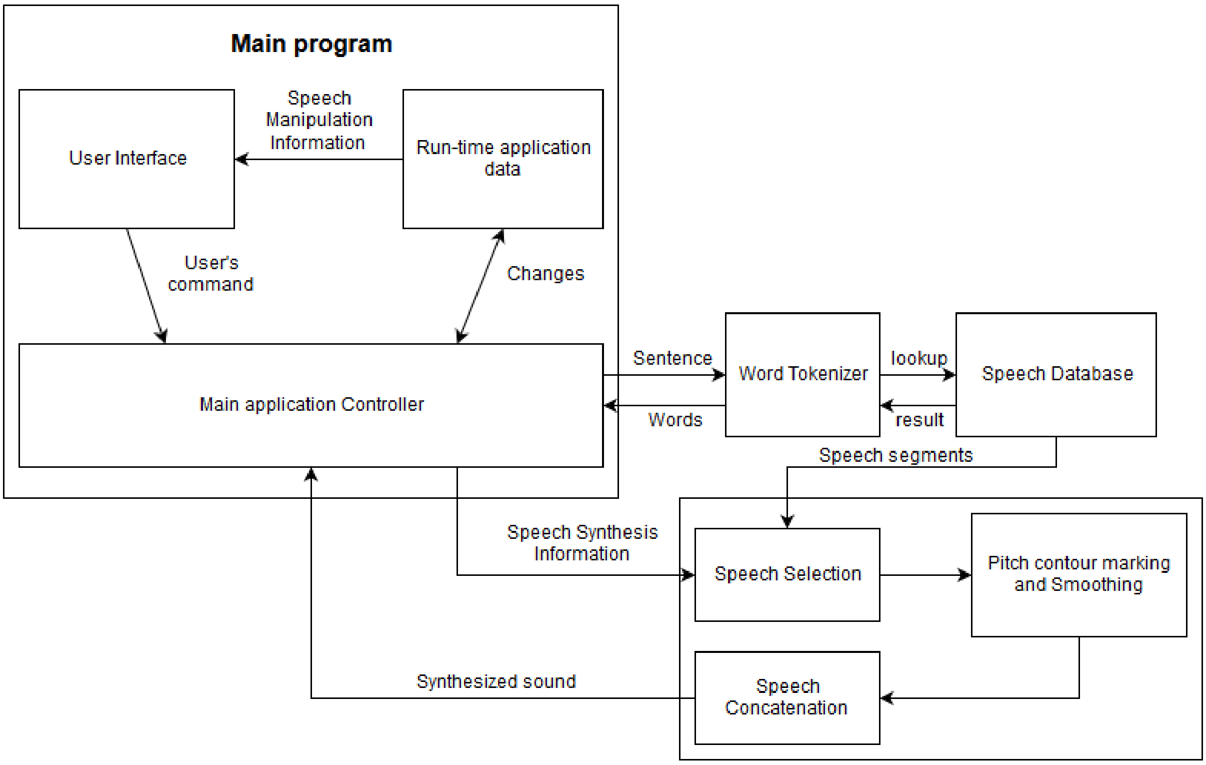


Figure 4 Software Architecture Diagram

Our project divided into 3 sections. First is front stage. We use Model-View Controller diagram to describe the frontend. In this component, there are 3 sub-component user-interface, main application controller and runtime model. The user-interface is a component for displaying the result to user and receiving the input from the user. The runtime model is a component to manage the progress in this system and show working state back to the user-interface. The controller diagram is a component for passing all the input to other main component and passing output from the other component back to the front stage.

Second main section is word tokenizer. The functionalities of this main component are finding the right or similar sentence from the speech database and concatenation each word to synthesis the selected speech that match the input text.

Final main section is speech manipulation component. This section functionalities are taking the speech segment sound file and using speech synthesis form controller and then marking pitch contour, duration and pitch modification. After all speech modifications, the modified speeches will be concatenated and forwarded to the main controller as a result.

3.4 Speech Database Design

Speeches database compose of sound files and metadata. Sound files metadata include sound file's name, duration, mean pitch level, corresponding word, preceding word in recording, and following word in recording. Sound files are stored in single folder, where metadata are stored in single-table flat-file database.

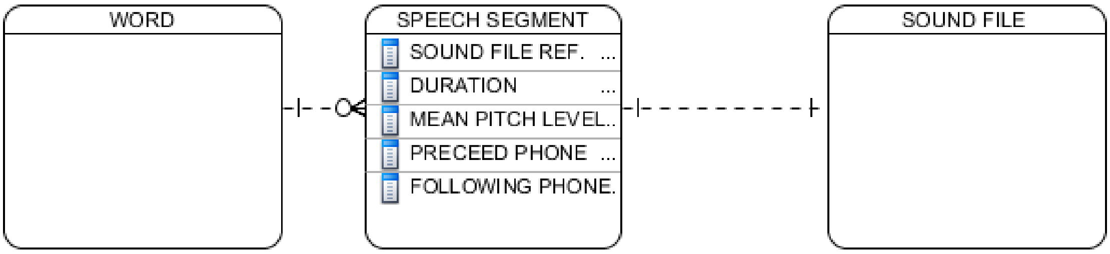


Figure 5 ER Diagram for syllable based speech database

The flat file database provide 2 interface to use. First interface is to find if certain syllable exists in database or not. Another interface is to pick nearest match speech in database, given desired context information

Various songs were recorded using single singer. The recording then will be segments into files, each corresponding to a syllable. Each recorded syllable then will be tagged with metadata and stored in the database.

3.5 Speech Selection Procedure

In the speech selection procedure, speech database will be looked up for desired syllable. Afterward context data were used to calculate a distance cost between desired speech and recorded speech. The context data are duration, mean pitch level, preceding syllable and following syllable. Recorded speech with lowest distance cost to desired speech will be selected.

3.6 Speech Manipulation

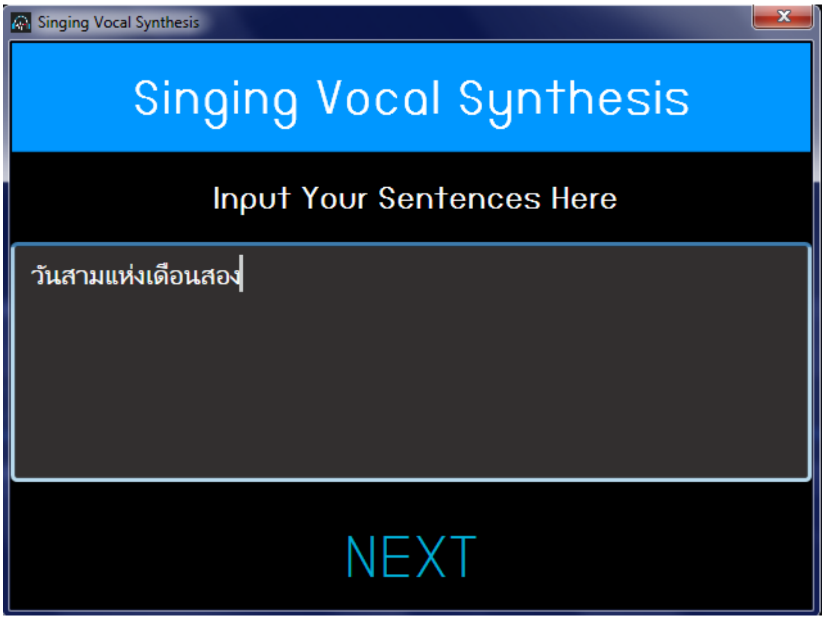
To give synthesized speech to a desired length with natural sounding, duration and pitch contour of recorded speech are manipulated using PSOLA algorithm in Praat software. Duration manipulation are quite straight forward as PSOLA algorithm use duration factor as input. As we know desired duration and recorded speech duration, the duration factor can be computed by

$$Duration\ factor = \frac{desired\ duration}{recorded\ speech\ duration}$$

Pitch manipulation are more complex as the pitch contour contain perceived tone of the word, which has meaning in Thai language. Using recorded speech's pitch contour and desired pitch level, desired pitch contour is generated using the following procedure

1. Extract recorded speech's pitch contour from each file.
2. Convert pitch contour from Hertz scale into semitone scale
3. Calculate median of pitch contour from each recorded speech
4. Calculate pitch difference (desired pitch – median of recorded speech)
5. Add pitch difference to all pitch point in recorded speech
6. Concatenate pitch points from multiple recorded speech together
7. Smoothen pitch contour using moving average
8. Convert pitch contour from semitone scale to Hertz scale

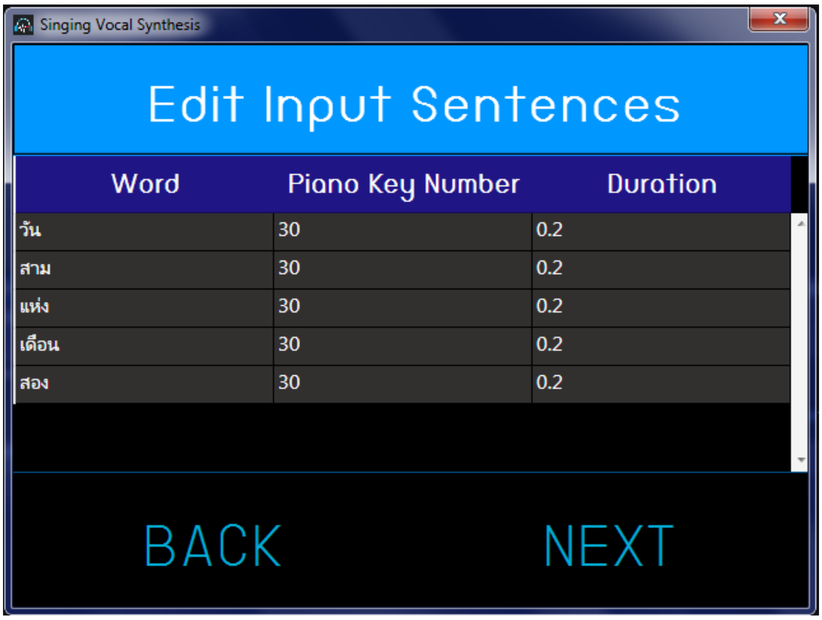
3.7 Software User Interface



วันสามแห่งเดือนสอง

Figure 6 User Interface Design: Input Scene

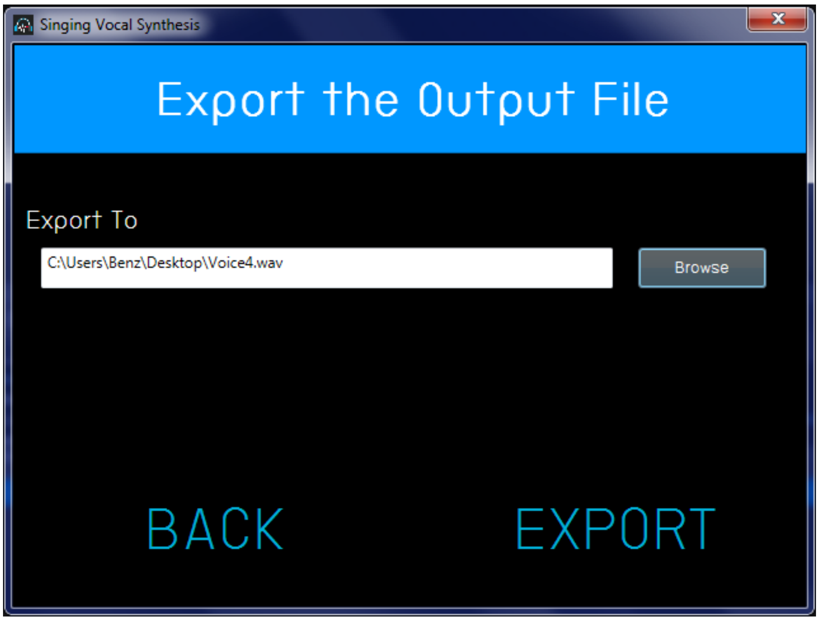
In this user interface, user can type the input as word in the textbox then click the NEXT button to continue to the next scene. Also the textbox will change to Thai language automatically so user can type immediately.



Word	Piano Key Number	Duration
วัน	30	0.2
สาม	30	0.2
แห่ง	30	0.2
เดือน	30	0.2
สอง	30	0.2

Figure 7 User Interface Design: Edit Value Scene

In this scene, User can adjust the tone by using piano key number which is referenced to the frequency that piano can producing in each note and octave. User can adjust the duration that corresponding in each word then user can go to the export scene by clicking NEXT button. User can edit the sentence by clicking BACK button to go back to the previous scene.



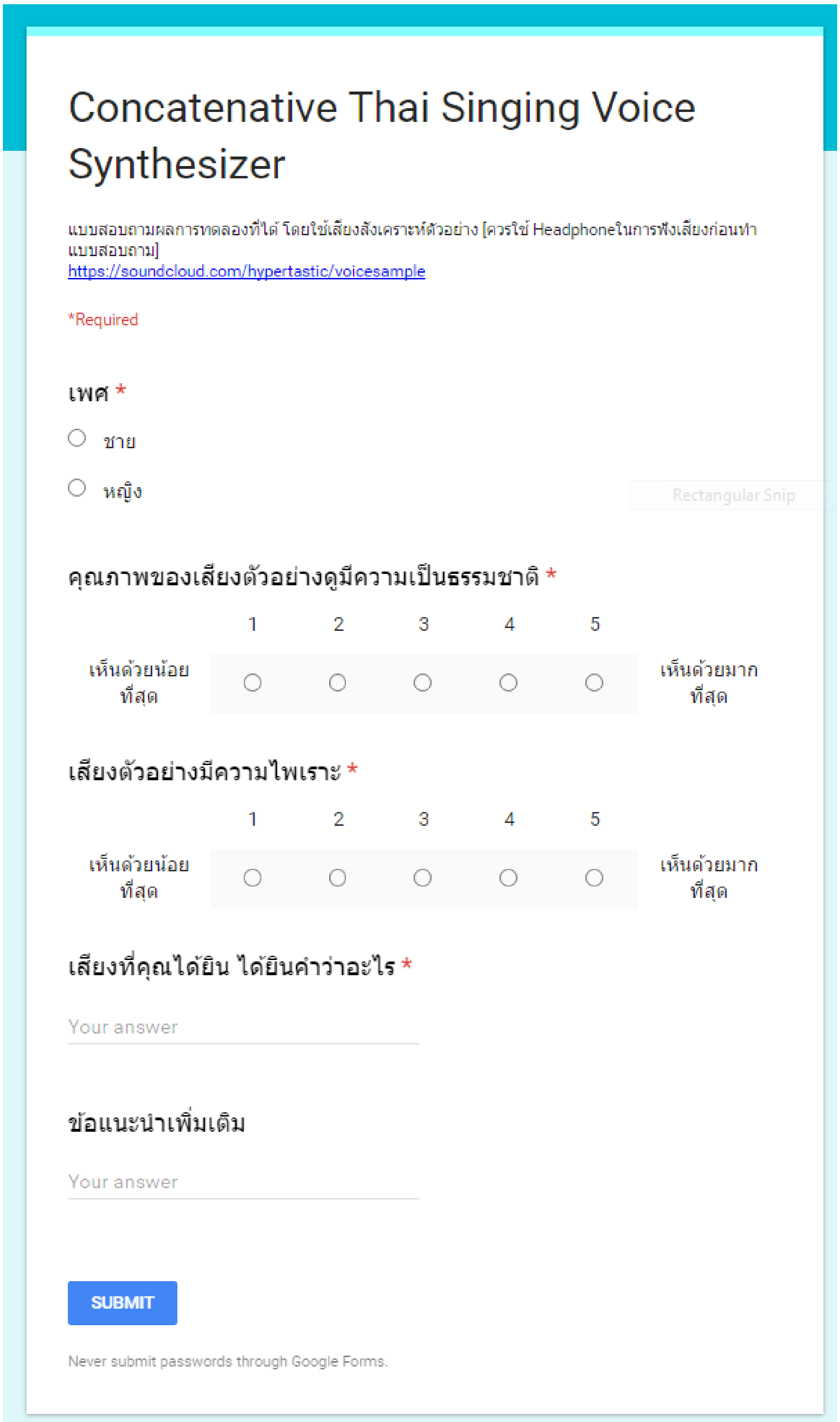
C:\Users\Benz\Desktop\Voice4.wav

Figure 8 User Interface Design: Export Scene

In export scene, User can select the directory of the output file by clicking Browse button then choose the destination folder and create the filename. After user choose the target of the output file, user can click the EXPORT button to export the file to the destination that user chose. The output file will be in Waveform Audio File Format.

This software will be made by using .NET C# Language as library programming language and Window Presentation Foundation as a platform for the user interface. Microsoft Virtual Studio and Blend are used for create and design the user interface.

3.8 Synthesized Vocal Evaluation



Concatenative Thai Singing Voice Synthesizer

แบบสอบถามผลการทดลองที่ได้ โดยใช้เสียงสังเคราะห์ตัวอย่าง [ควรใช้ Headphoneในการฟังเสียงก่อนทำแบบสอบถาม]
https://soundcloud.com/hypertastic/voicesample

*Required

เพศ *

○ ชาย

○ หญิง

คุณภาพของเสียงตัวอย่างดูมีความเป็นธรรมชาติ *

	1	2	3	4	5	
เห็นด้วยน้อยที่สุด	○	○	○	○	○	เห็นด้วยมากที่สุด

เสียงตัวอย่างมีความไพเราะ *

	1	2	3	4	5	
เห็นด้วยน้อยที่สุด	○	○	○	○	○	เห็นด้วยมากที่สุด

เสียงที่คุณได้ยิน ได้ยินคำว่าอะไร *

Your answer

ข้อแนะนำเพิ่มเติม

Your answer

Never submit passwords through Google Forms.

Figure 9 Evaluation Questionnaire

We export some of sample voices to create new voice file using this program and create questionnaire to ask people about the natural of voice and quality of voice.

3.9 List of Software used in the Project

- .Net Framework 4.5
- Visual Studio 2013
- Praat 6.0.14
- Java SE Runtime Environment 8
- Java SE Development Kit 8
- SqliteBrowser 3.8.0

Chapter 4
Results and Discussion

4.1 Difference between Normal Voice and Singing Voice

The comparison between normal speaking voice and singing voice gives two interesting results as shown below

1. **Pitch level and voice duration**

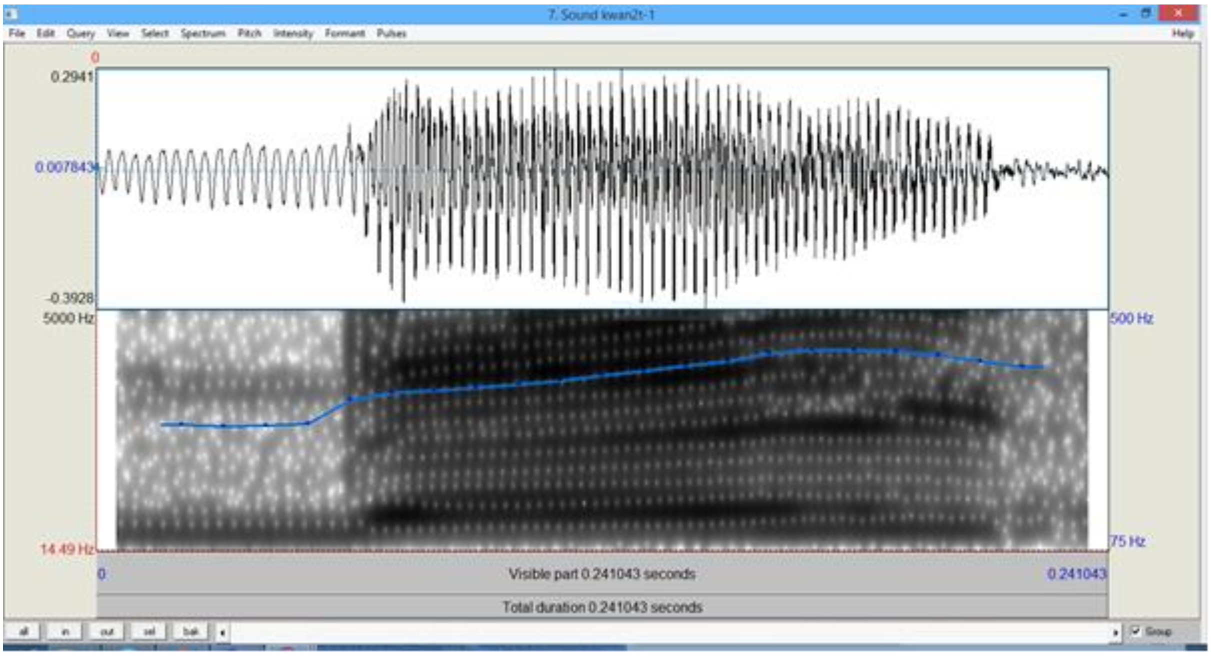

7. Sound kwan2t-1
File Edit Query View Select Spectrum Pitch Intensity Formant Pulses
Help
0
0.2941
0.007843
-0.3928
5000 Hz
14.49 Hz
500 Hz
75 Hz
0
Visible part 0.241043 seconds
0.241043
Total duration 0.241043 seconds
all in out sel bak
Group

Figure 10 Normal speaking voice of word“Dai”(ได้) located at beginning of the sentence

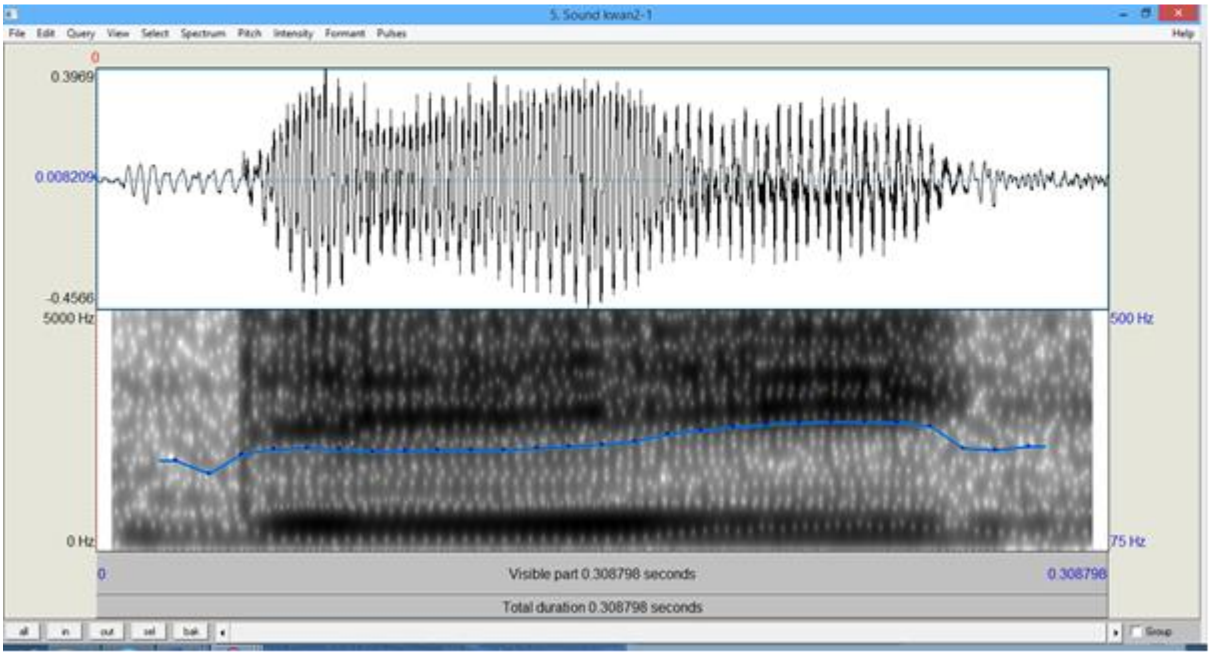

5. Sound kwan2-1
File Edit Query View Select Spectrum Pitch Intensity Formant Pulses
Help
0
0.3969
0.008209
-0.4566
5000 Hz
0 Hz
500 Hz
75 Hz
0
Visible part 0.308798 seconds
0.308798
Total duration 0.308798 seconds
all in out sel bak
Group

Figure 11 Singing voice of word“Dai”(ได้) located at beginning of the sentence

The word “Dai” shown above is first pair of interesting results, with wave from, spectrogram and pitch contour shown. The word “Dai” have falling tones, according to Thai language’s five phonemic tones. It is shown that both singing voice and speaking voice have same shape of pitch contour (drawn as blue curve lines). However, there is visible difference of pitch level in pitch contour. So from this result, two properties of voice that should be customizable are found, which are pitch level and voice duration.

2. **Effect of singing on pitch contour**

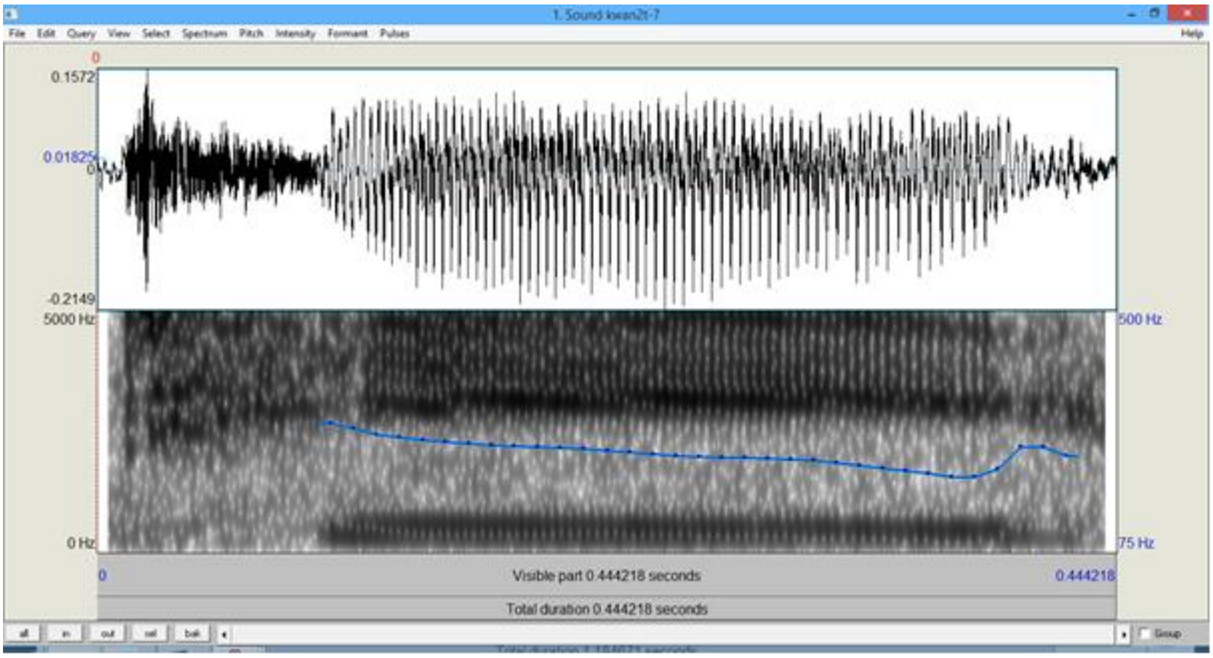

1. Sound kwan2t-7
File Edit Query View Select Spectrum Pitch Intensity Formant Pulses
Help
0
0.1572
0.01825
0
-0.2149
5000 Hz
0 Hz
500 Hz
75 Hz
0
Visible part 0.444218 seconds
0.444218
Total duration 0.444218 seconds
all in out sel bak
Group

Figure 12 Normal speaking of word“Tee”(ที่)

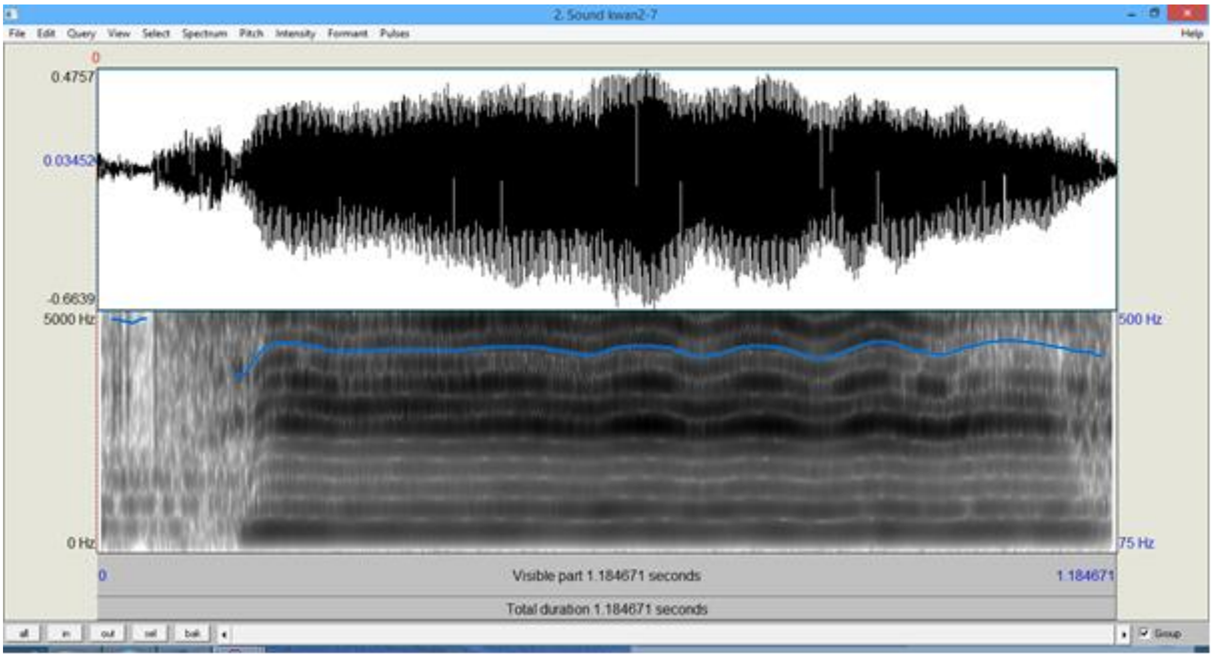

2. Sound kwan2-7
File Edit Query View Select Spectrum Pitch Intensity Formant Pulses
Help
0
0.4757
0.03452
-0.6639
5000 Hz
0 Hz
500 Hz
75 Hz
0
Visible part 1.184671 seconds
1.184671
Total duration 1.184671 seconds
all in out sel bak
Group

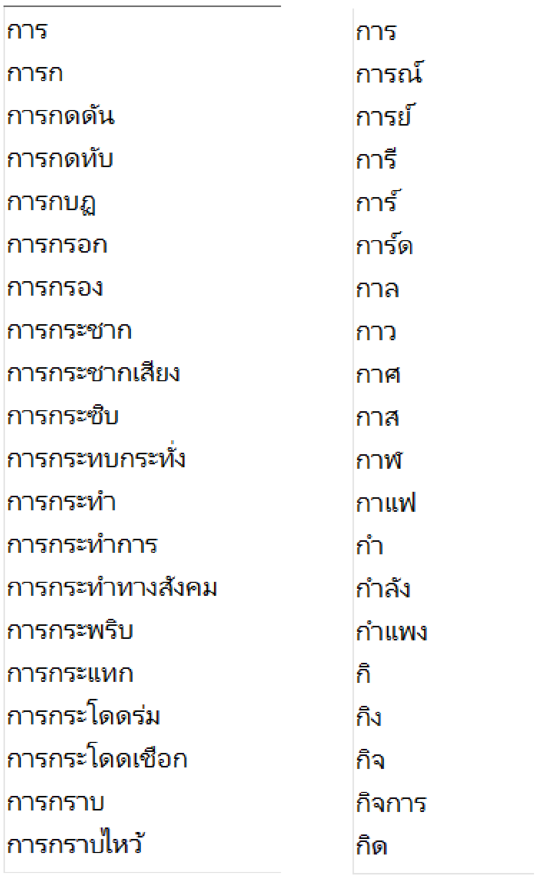
Figure 13 Singing voice of word“Tee”(ที่)

The word “Tee” with falling tone is second interesting pair of results. There is overshoot, which is a sudden rise in pitch, at the early beginning of the singing voice. Also there is vibrato, which is pulsating change of pitch, along the singing voice. These two musical effect help singing voice to be perceived as more lively sound. Also singing voice tends to have these two musical effect than normal speaking voice. So these two musical effect should be included in the vocal synthesis software for song as well.

According to our small experiment, when speaking the vocal muscles and diaphragm are less controlled. They can fluctuate over the same notes as when singing, they just do not normally do that because it's distracting to the listener. It is possible to do vibrato and other singing mechanics while speaking it's just not useful. It's all about breath control and which set of vocal muscles are in better shape. People tend to develop whatever already works better.

4.2 Word Tokenizer and Dictionary

Word tokenization are based on NECTEC’s Lexto program, which use dictionary of words to tokenize. Normally Lexto have Lexitron as default dictionary, in which input text are split into words. However, our software required text to be tokenized by syllables. So a custom dictionary was made by splitting words in Lexitron into syllables. Parts of dictionary are shown below



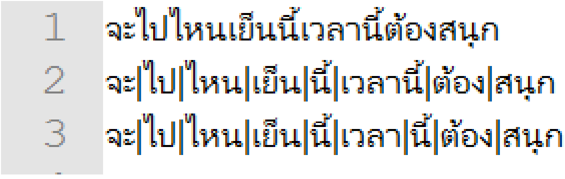
การ
การก
การกดดัน
การกดทับ
การกบฏ
การกรอก
การกรอง
การกระชาก
การกระชากเสียง
การกระซิบ
การกระทบกระทั่ง
การกระทำ
การกระทำการ
การกระทำทางสังคม
การกระพริบ
การกระแทก
การกระโดดร่ม
การกระโดดเชือก
การกราบ
การกราบไหว้

การ
การณ์
การย์
การี
การ์
การ์ด
กาล
กาว
กาศ
กาส
กาฬ
กาแฟ
กำ
กำลัง
กำแพง
กิ
กิง
กิจ
กิจการ
กิด

Figure 14 Some parts of dictionary file

As figure 13 shown that, on the left column and right column are from Lexitron dictionary and custom dictionary respectively.

The example of word tokenization are shown below



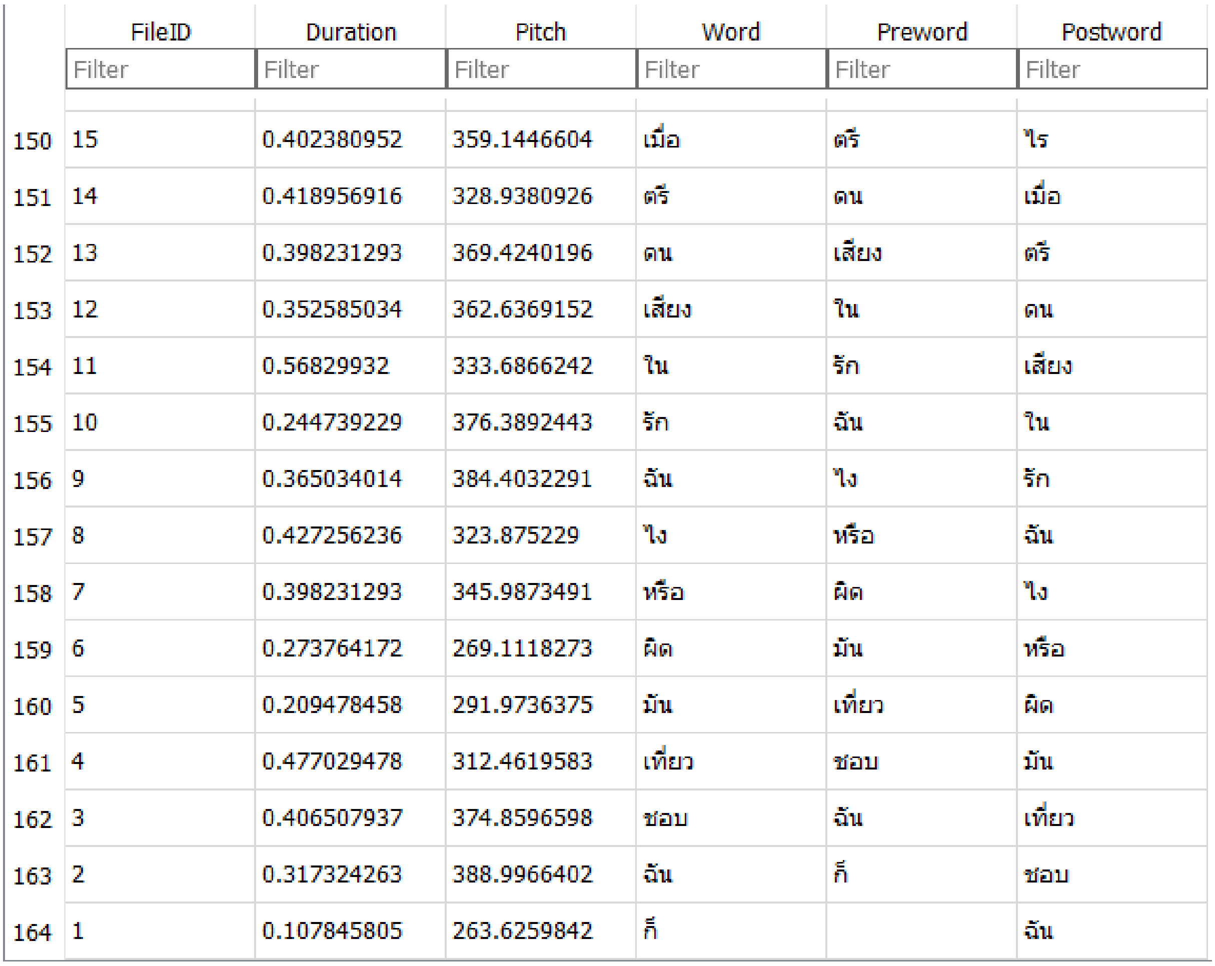
```
จะไปไหนเย็นนี้เวลานี้ต้องสนุก
จะ|ไป|ไหน|เย็น|นี้|เวลานี้|ต้อง|สนุก
จะ|ไป|ไหน|เย็น|นี้|เวลา|นี้|ต้อง|สนุก
```

Figure 15 The differences of using different dictionaries

The first line, second line and third line are the input, the word tokenization using Lexitron dictionary and the word tokenization using custom dictionary respectively

4.3 Concatenative Speech Database and Lookup Procedure

Speech database use a directory of sound files to store the recorded sound, with a SQLite database file to contain metadata in a single table. Part of database file, viewed via SqliteBrowser is shown below.



	FileID	Duration	Pitch	Word	Preword	Postword
	Filter	Filter	Filter	Filter	Filter	Filter
150	15	0.402380952	359.1446604	เมื่อ	ตรี	ไร
151	14	0.418956916	328.9380926	ตรี	ดน	เมื่อ
152	13	0.398231293	369.4240196	ดน	เสียง	ตรี
153	12	0.352585034	362.6369152	เสียง	ใน	ดน
154	11	0.56829932	333.6866242	ใน	รัก	เสียง
155	10	0.244739229	376.3892443	รัก	ฉัน	ใน
156	9	0.365034014	384.4032291	ฉัน	ไง	รัก
157	8	0.427256236	323.875229	ไง	หรือ	ฉัน
158	7	0.398231293	345.9873491	หรือ	ผิด	ไง
159	6	0.273764172	269.1118273	ผิด	มัน	หรือ
160	5	0.209478458	291.9736375	มัน	เที่ยว	ผิด
161	4	0.477029478	312.4619583	เที่ยว	ชอบ	มัน
162	3	0.406507937	374.8596598	ชอบ	ฉัน	เที่ยว
163	2	0.317324263	388.9966402	ฉัน	ก็	ชอบ
164	1	0.107845805	263.6259842	ก็		ฉัน

Figure 16 Database containing recorded speech metadata

Despite FileID is the primary key of the table, word column is looked up first for matched word. If the desired word to be synthesized have preceding word or following word match the recorded segment in database, the recorded segment will gain score. The recorded segment with similar duration and pitch also gain score based on similarity. Word with highest score are picked to be used for synthesis.

4.4 Implemented User Interface

As we designed the user interface prototype, the interface is too complicated to understand how to generate the tokenized word into the grid. So we overhauled the old user interface into the new user interface by using Windows Presentation Foundation as a platform and .NET framework as a library. We also used Blend to design the interface and Virtual Studio to design the controller and interface mechanism. The result is that the new interface is simpler and easier to use than the old version.

4.5 Evaluation Result

Five surveys with the synthesized voice were made to find factors that affect the voice quality. The survey asked participants to evaluate the voices naturalness and melodiousness. High naturalness voice means that the voice is similar to human voice. And high melodiousness voice means that the voice is suitable for singing.

For each survey, the independent variable are one of the following

- Recorded voice source
- Synthesized voice pitch
- Synthesized voice pitch under same phrase context
- Synthesized voice duration
- Synthesized voice duration under phrase context

The survey result can be found in appendix.

4.5.1 Evaluation of Voice with Different Recorded Voice Source

There are 2 voice samples in this set. The first voice sample was generated from singing recorded voice. And the second voice sample was generated from speaking recorded voice. Survey result shows that second sample have higher naturalness and melodiousness than the first sample. Some participants says that the first voice sample are not listenable.

Sample	1	2
Recorded Voice Source	Singing Voice	Speaking Voice
Naturalness	2.58	3.73
Melodiousness	3.42	3.80

Table 6 Servey Result of Different Recorded Voice Source

4.5.2 Evaluation of Voice with Different Voice Pitch

There are 5 voice samples in this set. Every voice speak a single word with different pitch level. The recorded word that used have pitch level 37.8.

Sample	1	2	3	4	5
Pitch level (piano key number reference)	30	36	42	48	54
Pitch difference (semitone)	-7.8	-1.8	+4.2	+10.2	+16.2
Naturalness	2.73	3.60	3.67	2.87	2.40
Melodiousness	2.67	3.47	3.60	3.07	2.53

Table 7 Servey Result of Different Voice Pitch

The synthesized voice that have pitch higher than the recorded voice's pitch by small amount have higher naturalness and melodiousness.

4.5.3 Evaluation of Voice with Different Voice Pitch under Same Phrase Context

There are 5 voice samples in this set. Every voice speak a three-word phrase. The first and the last words have constant pitch level at 44 and 38 respectively. These value are picked from the original recorded voice's pitch level, so the modification of first and last word is minimal. The middle word have pitch level varied.

Sample	1	2	3	4	5
Middle word's pitch level (piano key difference)	30	36	42	48	54
Sum of pitch difference to adjacent words (semitone)	-22	-10	-4	+14	+26
Naturalness	2.07	2.60	3.60	2.87	2.27
Melodiousness	2.07	2.67	3.47	3.00	2.53

Table 8 Servey Result of Different Voice Pitch under Same Phrase Context

The synthesized voice with low pitch difference to adjacent words gives high naturalness and melodiousness. However, for the middle word, effect of pitch difference between synthesized voice and recorded voice might be present but not concerned.

4.5.4 Evaluation of Voice with Different Voice Duration

There are 5 voice samples in this set. Every voice speak a single word with varied duration. Duration multiplier is the independent variable. Originally, the recorded voice have duration of 0.238 second.

Sample	1	2	3	4	5
Voice's duration (second)	0.2	0.3	0.4	0.5	0.6
Duration multiplier from the original voice	0.84	1.26	1.68	2.10	2.52
Naturalness	3.33	3.93	4	4.07	3.53
Melodiousness	3.67	3.80	3.87	3.60	3.47

Table 9 Survey Result of Voice with Different Voice Duration

Participants says that voice duration of 0.4 to 0.5 have high naturalness and melodiousness. Too long or too short voice duration impaired the naturalness and melodiousness.

4.5.5 Evaluation of Voice with Different Voice Duration under Same Phrase Context

There are 4 voice samples in this set. Every voice speak a three-word phrase. The first and the last words have duration fixed, while the middle word have duration varied. First word have duration 0.2 seconds, while the last word has duration of 0.4 seconds

Sample	1	2	3	4
Middle voice's duration (second)	0.1	0.2	0.3	0.4
Naturalness	4.33	4.27	3.33	2.60
Melodiousness	4.27	4.27	3.53	3.20

Table 10 Survey Result of Voice with Different Voice Duration under Same Phrase Context

When middle word has shorter duration compared to adjacent words, the naturalness and melodiousness is high.

4.6 Evaluation Result Descriptive Analysis

Singing recorded voice were used as voice database at first because we assume that recorded singing voice are useful to synthesize another singing voice. The assumption, however, is incorrect. During the survey, we found out that synthesized song have words' tone changed compared to the tone written in input text. So the desired tone to be synthesized should be added to synthesizer's parameter. Also using speaking voice as voice database give advantages compared to using singing voice as voice database. Firstly, the recorded voice are more general to be used to synthesize another voice. Secondly, by using speaking voice records, we can generate larger amount of recordings samples to be used as voice database. With higher amount of recording samples, higher quality voice can be synthesized.

From the evaluation surveys, it is shown that voice with low modification from the voice database gives more naturalness and melodiousness. However, by modifying pitch and duration of some word can give higher naturalness when the word appear with another words.

To implement a synthesizer system for singing with this approach, it is more practical to record speaking voice as voice database and then add singing temperament as needed later. As far as we know, desired tone should be another adjustable parameter along with duration and pitch level.

Chapter 5
Conclusion

5.1 Progress

Task	Status
Software Requirement Analysis	Completed
Architectural Design	Completed
User Interface	Completed
Main application Controller	Completed
Run-time application data	Completed
Word Tokenizer	Completed
Speech Database	Small set (164 voice files) recorded for demonstration.
Speech Database Unit selection	Completed
Pitch contour marking and smoothing	Completed
Speech concatenation	Completed
Software and Result evaluation design	Completed
Software and Result evaluation	Completed

Table 11 Progress Table

5.2 Problem & Solution

The first problem comes from ambiguous requirement. At first we desired to develop a speech synthesis system without clear description about what software can do, how user interact with software and how the speech database would be designed. Also programming experience of developers are limited, thus large amount of time are needed to learn the programming language, technology and framework. So software development experience takes major roles in project progress. The second problem is about the speech database recording. The amount of word recorded is large and speech database generation cannot be automated. We decided to work on this task earlier. The third problem is the environment noise that can affect the recording. The noise are avoided by finding the noiseless environment for recoding and used appropriate tools in recording. The last problem is Thai language text encoding. A third party software we used processes text in TIS-620 encoding, while other software process text in UTF-8 encoding. Short term solution is to keep track on encoding of files along the process. While possible long term solution is to rewrite the open source software in desired environment, thus we can control file encoding used.

References

T. Dutoit and H. Leich. *“MBR-PSOLA: Text-To-Speech Synthesis Based on an MBE Re-synthesis of the Segments Database”*. (1993)

E. Moulines and F. Charpentier. *“Pitch-Synchronous Waveform Processing Techniques For Text-to-speech Synthesis Using Diphones”*. (1990)

H. Valbret, E. Moulines and J.P. Tubach. *”Voice Transformation Using PSOLA Technique”*. (1992)

A.j. Hunt and A.w. Black. *“Unit Selection in a Concatenative Speech Synthesis System Using a Large Speech Database”*. (1996)

Appendix

Full Survey Result

Each voice sample has the same question set.

Question 1: Naturalness of synthesized voice

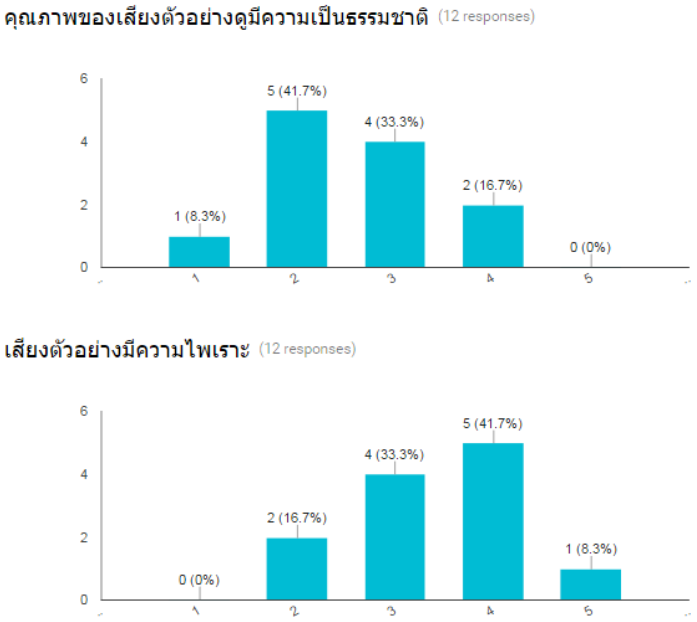
(Linear scale from 1 to 5 with 1 is not natural and 5 is natural)

Question 2: Melodiousness of synthesized voice

(Linear scale from 1 to 5 with 1 is bad and 5 is good)

Evaluation of Voice with Different Recorded Voice Source

- Singing Voice as Voice Source (12 responses)



คุณภาพของเสียงตัวอย่างดูมีความเป็นธรรมชาติ (12 responses)

เสียงตัวอย่างมีความไพเราะ (12 responses)

Figure 17 Result Graph of Singing Voice as Database

- Speaking Voice as Voice Source (14 responses)

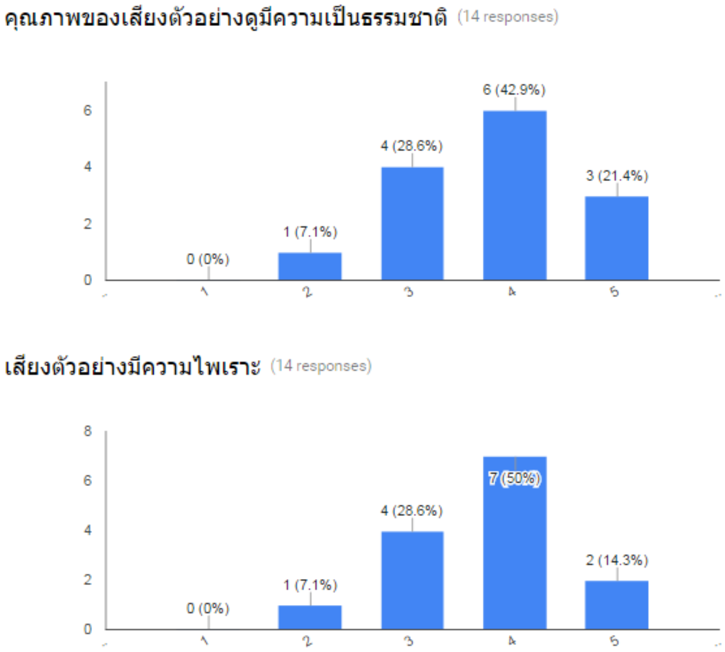
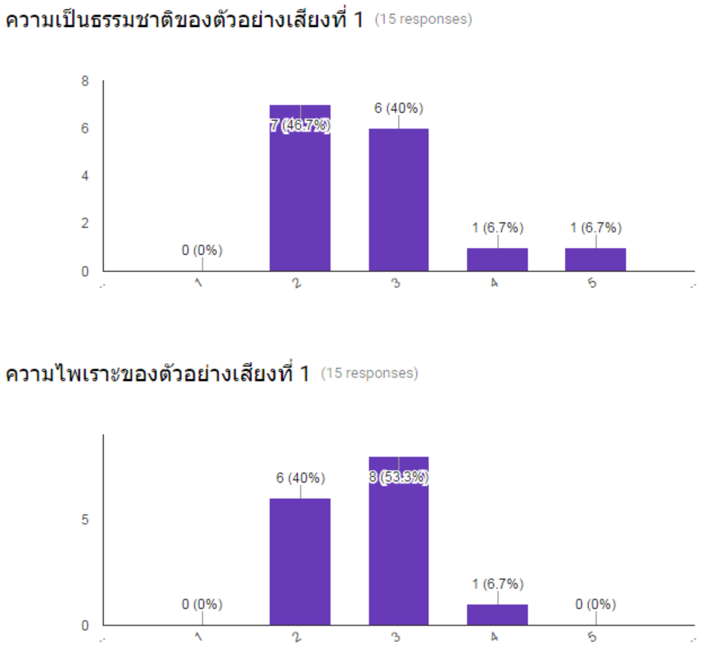


Figure 18 Result Graph of Speaking Voice as Database

Evaluation of Voice with Different Voice Pitch

- Sample 1

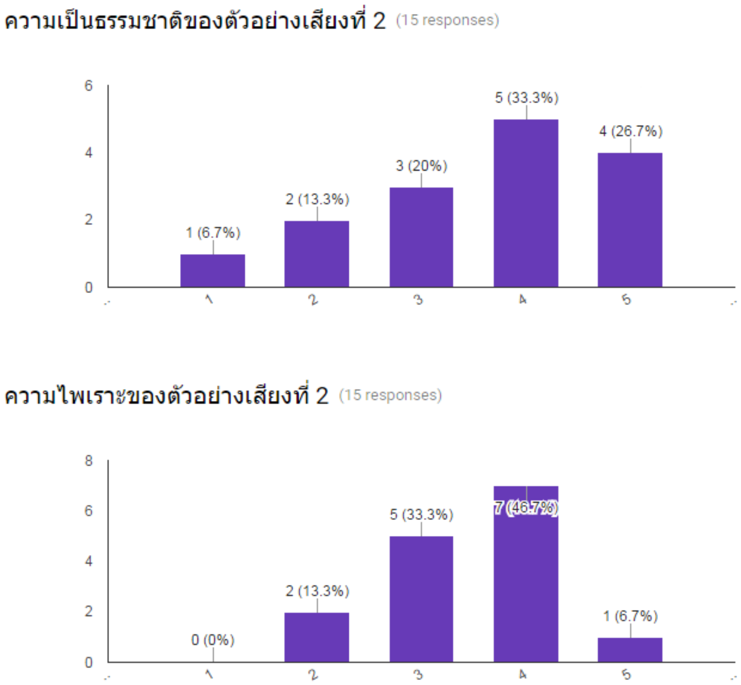


ความเป็นธรรมชาติของตัวอย่างเสียงที่ 1 (15 responses)

ความไพเราะของตัวอย่างเสียงที่ 1 (15 responses)

Figure 19 Result Graph of Sample 1 in Different Voice Pitch

- Sample 2

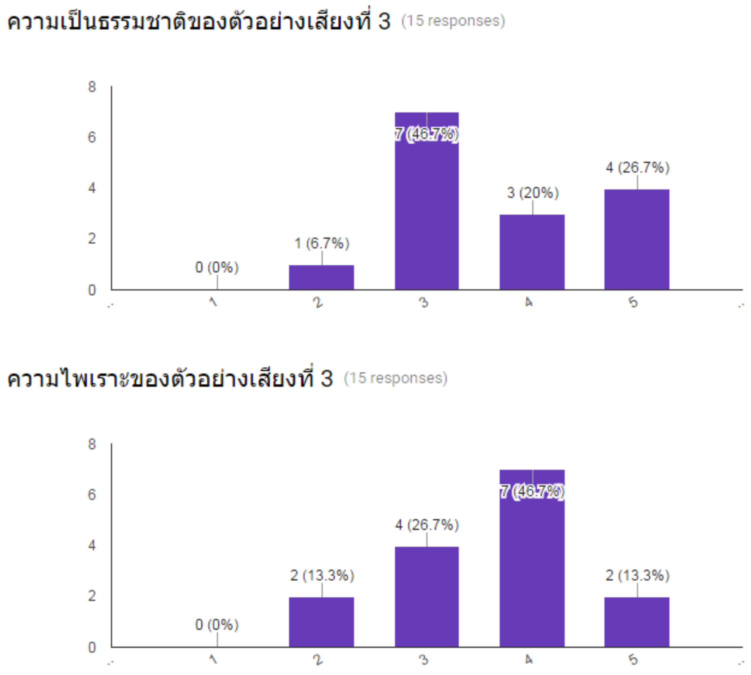


ความเป็นธรรมชาติของตัวอย่างเสียงที่ 2 (15 responses)

ความไพเราะของตัวอย่างเสียงที่ 2 (15 responses)

Figure 20 Result Graph of Sample 2 in Different Voice Pitch

- Sample 3

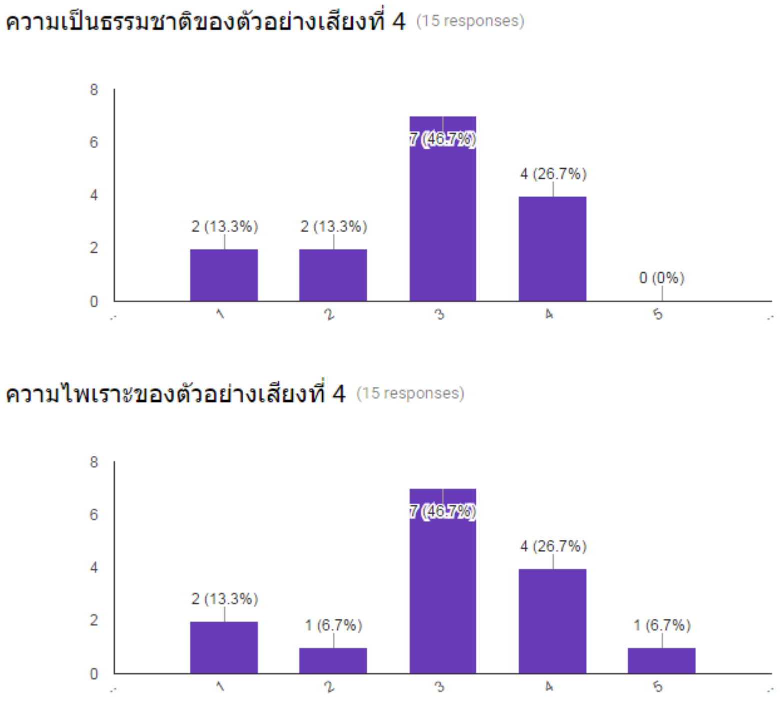


ความเป็นธรรมชาติของตัวอย่างเสียงที่ 3 (15 responses)

ความไพเราะของตัวอย่างเสียงที่ 3 (15 responses)

Figure 21 Result Graph of Sample 3 in Different Voice Pitch

- Sample 4



ความเป็นธรรมชาติของตัวอย่างเสียงที่ 4 (15 responses)

ความไพเราะของตัวอย่างเสียงที่ 4 (15 responses)

Figure 22 Result Graph of Sample 4 in Different Voice Pitch

- Sample 5

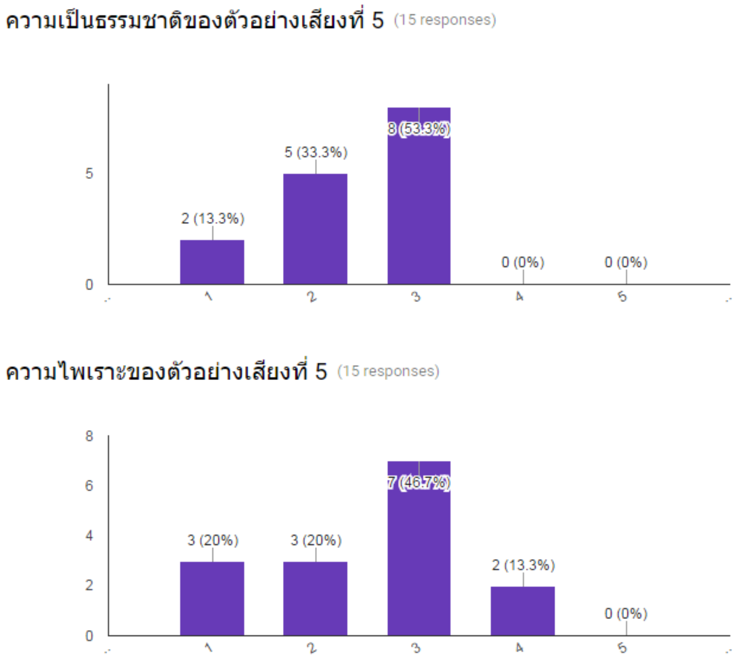
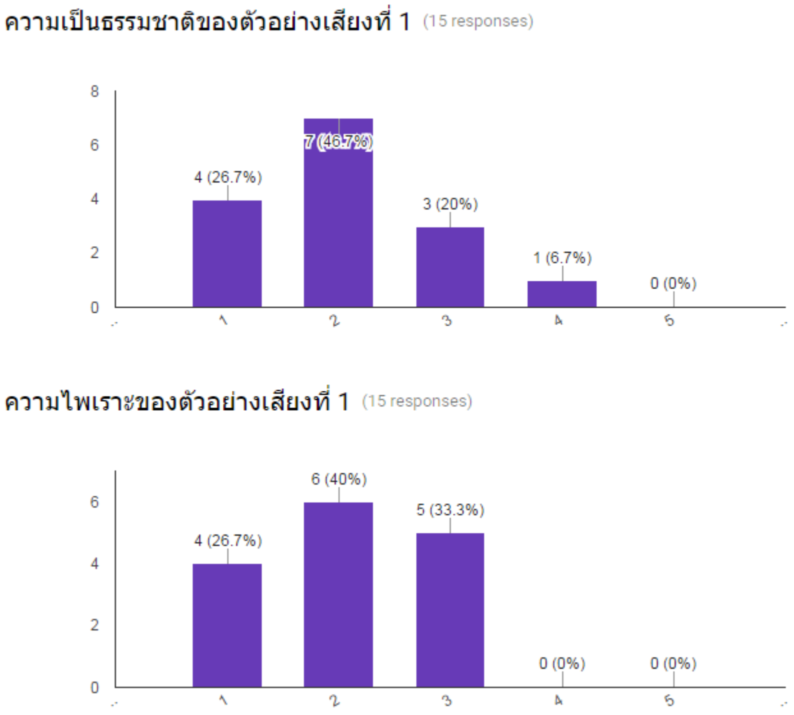


Figure 23 Result Graph of Sample 5 in Different Voice Pitch

Evaluation of Voice with Different Voice Pitch under Same Phrase Context

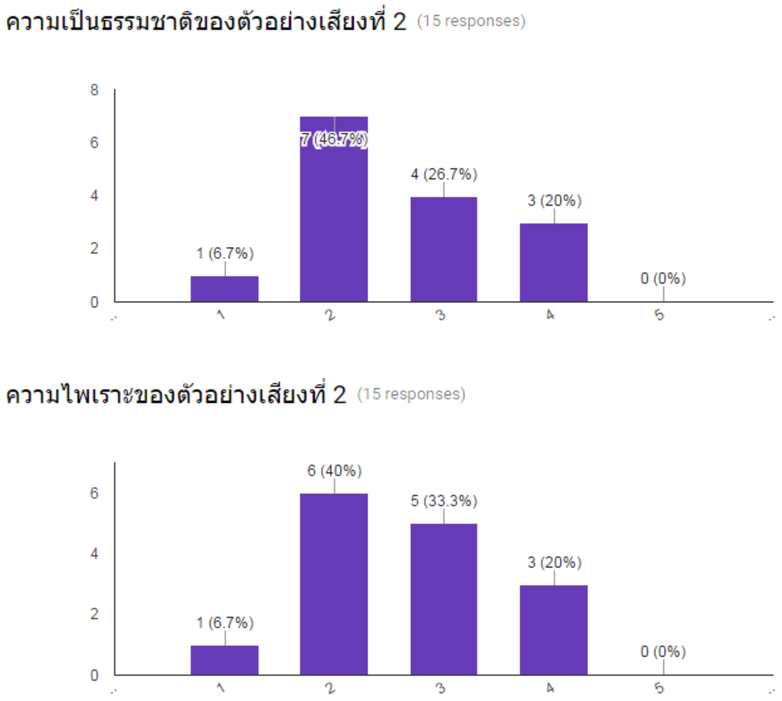
- Sample 1


ความเป็นธรรมชาติของตัวอย่างเสียงที่ 1 (15 responses)

ความไพเราะของตัวอย่างเสียงที่ 1 (15 responses)

Figure 24 Result Graph of Sample 1 in Different Voice Pitch under Same Phrase Context

- Sample 2


ความเป็นธรรมชาติของตัวอย่างเสียงที่ 2 (15 responses)

ความไพเราะของตัวอย่างเสียงที่ 2 (15 responses)

Figure 25 Result Graph of Sample 2 in Different Voice Pitch under Same Phrase Context

- Sample 3

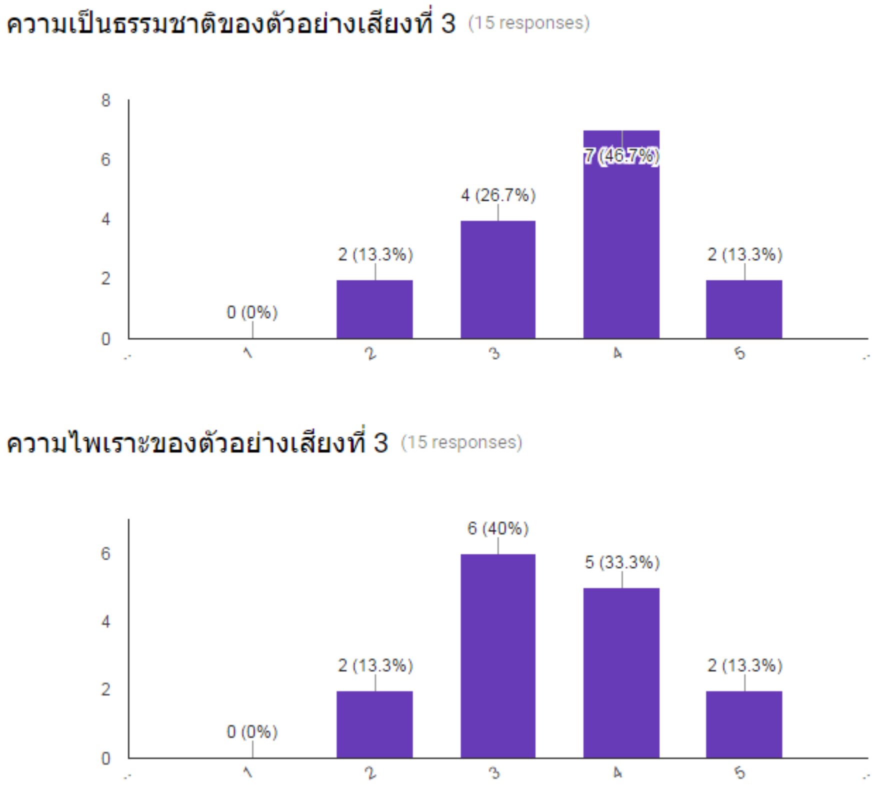


Figure 26 Result Graph of Sample 3 in Different Voice Pitch under Same Phrase Context

- Sample 4

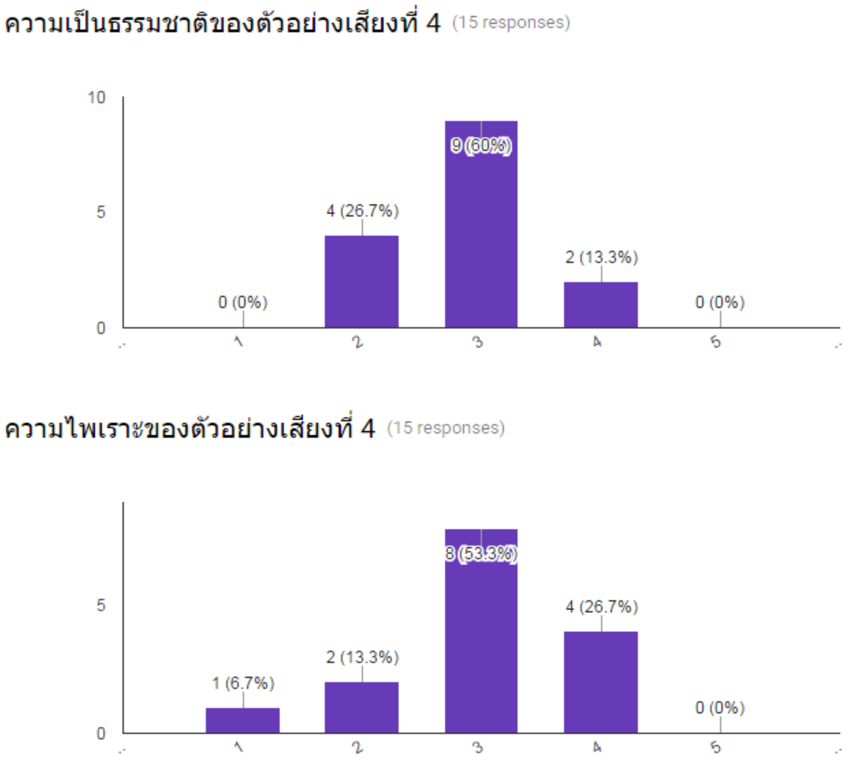


Figure 27 Result Graph of Sample 4 in Different Voice Pitch under Same Phrase Context

- Sample 5

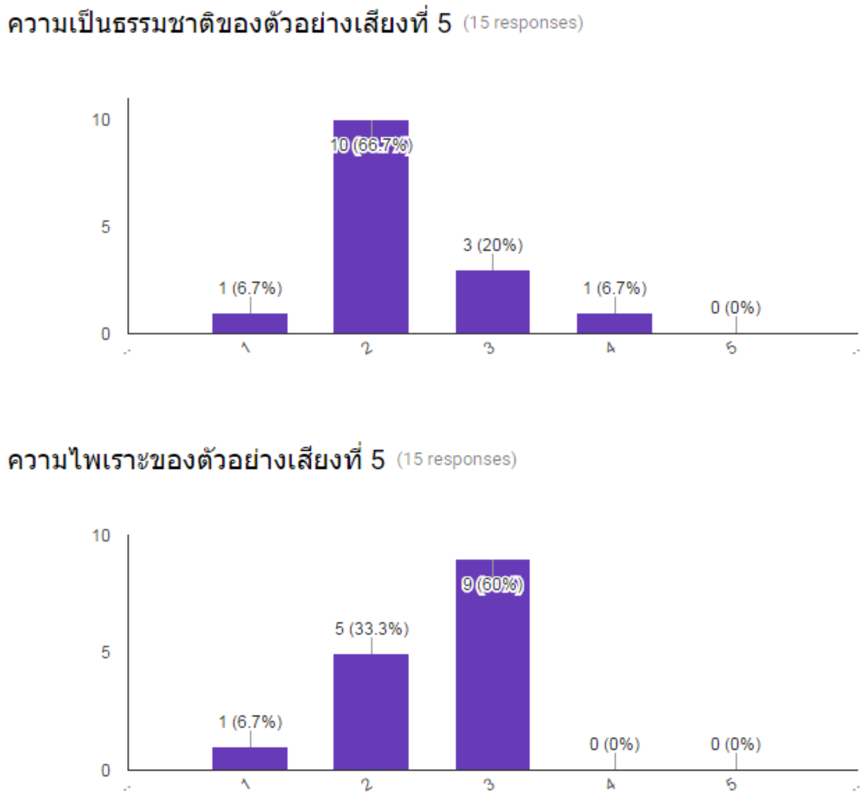


Figure 28 Result Graph of Sample 5 in Different Voice Pitch under Same Phrase Context

Evaluation of Voice with Different Voice Duration

- Sample 1

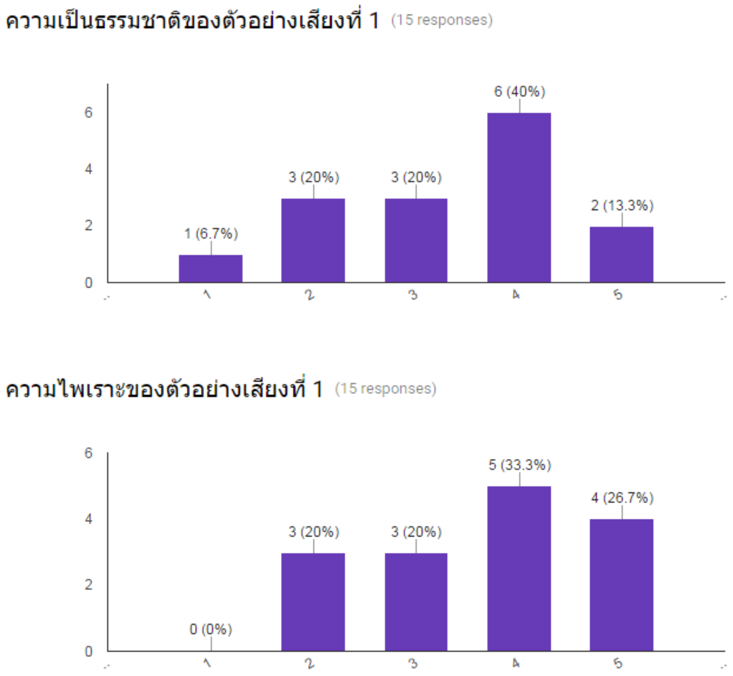


Figure 29 Result Graph of Sample 1 in Different Voice Duration

- Sample 2

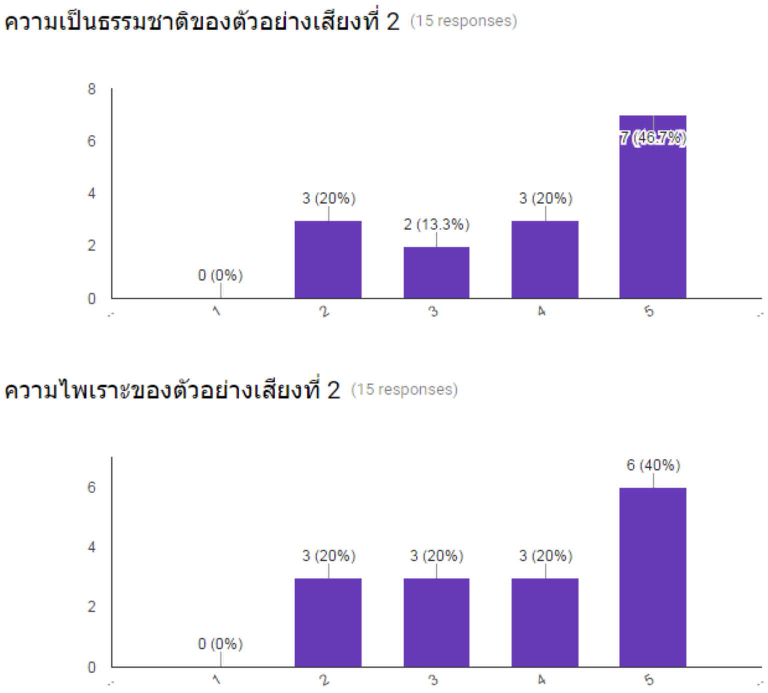
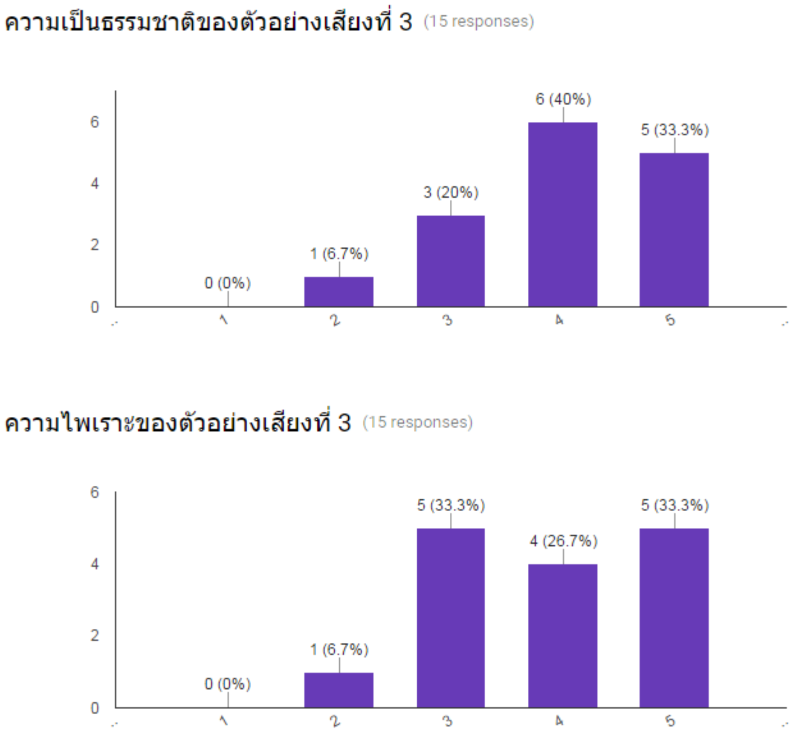


Figure 30 Result Graph of Sample 2 in Different Voice Duration

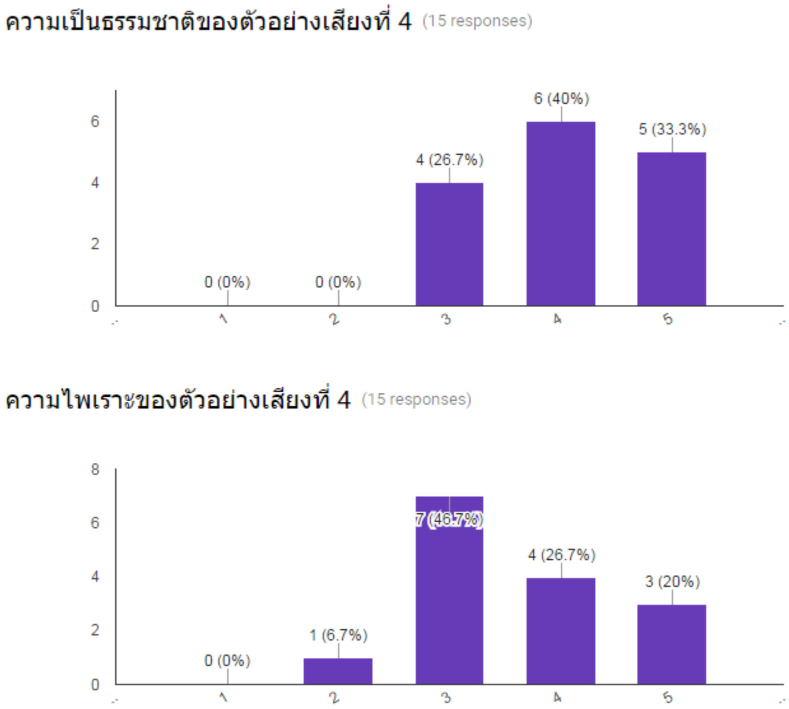
- Sample 3


ความเป็นธรรมชาติของตัวอย่างเสียงที่ 3 (15 responses)

ความไพเราะของตัวอย่างเสียงที่ 3 (15 responses)

Figure 31 Result Graph of Sample 3 in Different Voice Duration

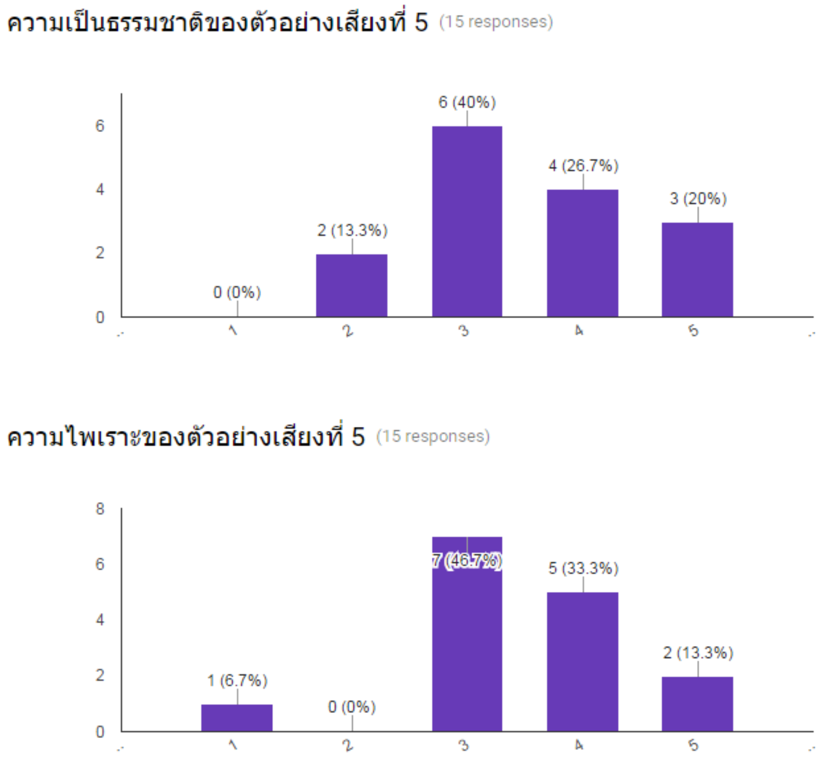
- Sample 4


ความเป็นธรรมชาติของตัวอย่างเสียงที่ 4 (15 responses)

ความไพเราะของตัวอย่างเสียงที่ 4 (15 responses)

Figure 32 Result Graph of Sample 4 in Different Voice Duration

- Sample 5



ความเป็นธรรมชาติของตัวอย่างเสียงที่ 5 (15 responses)

ความไพเราะของตัวอย่างเสียงที่ 5 (15 responses)

Figure 33 Result Graph of Sample 5 in Different Voice Duration

Evaluation of Voice with Different Voice Duration under Same Phrase Context

- Sample 1

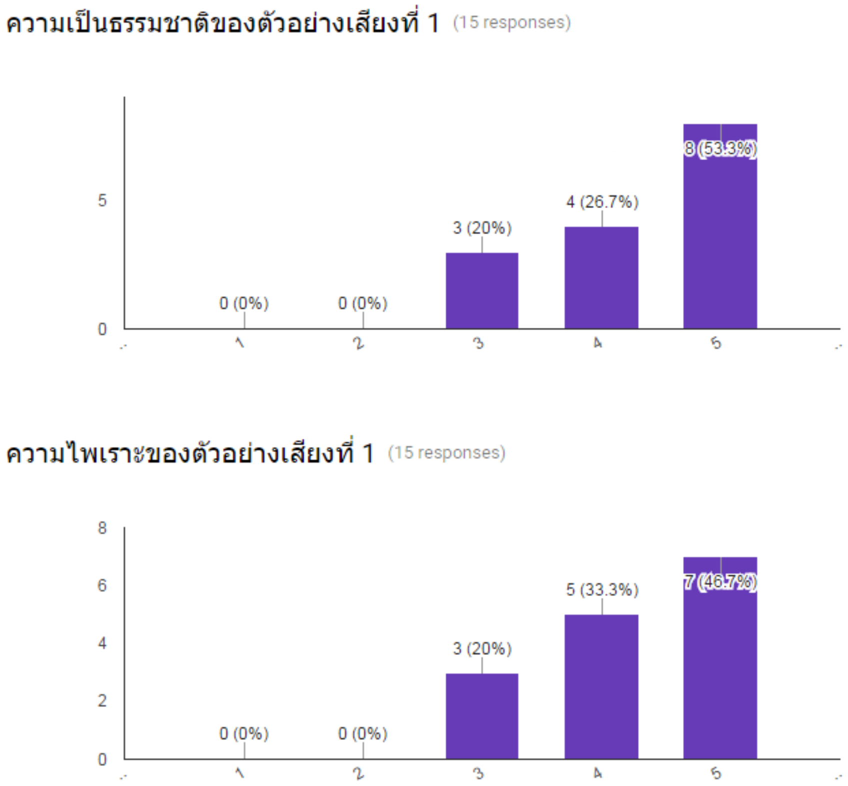


Figure 34 Result Graph of Sample 1 in Different Voice Duration under Same Phrase Context

- Sample 2

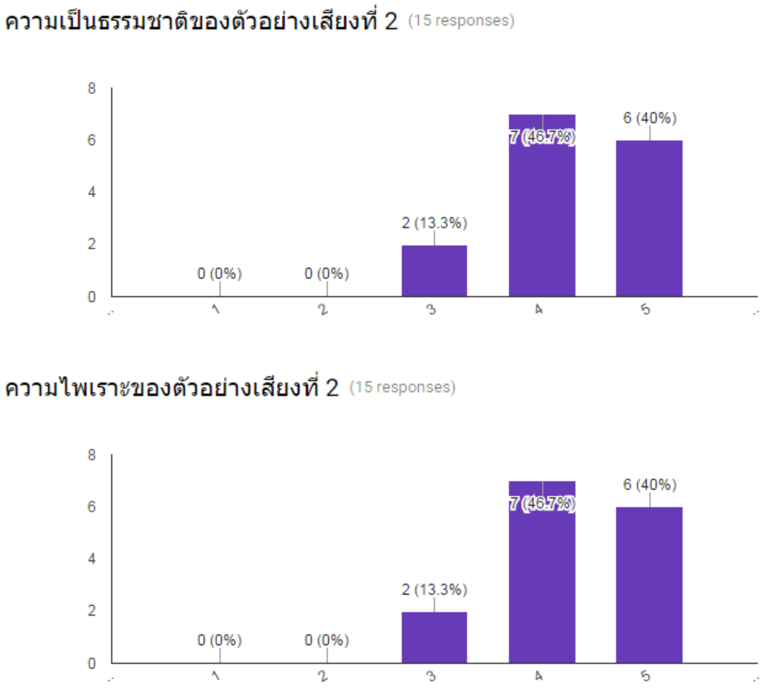
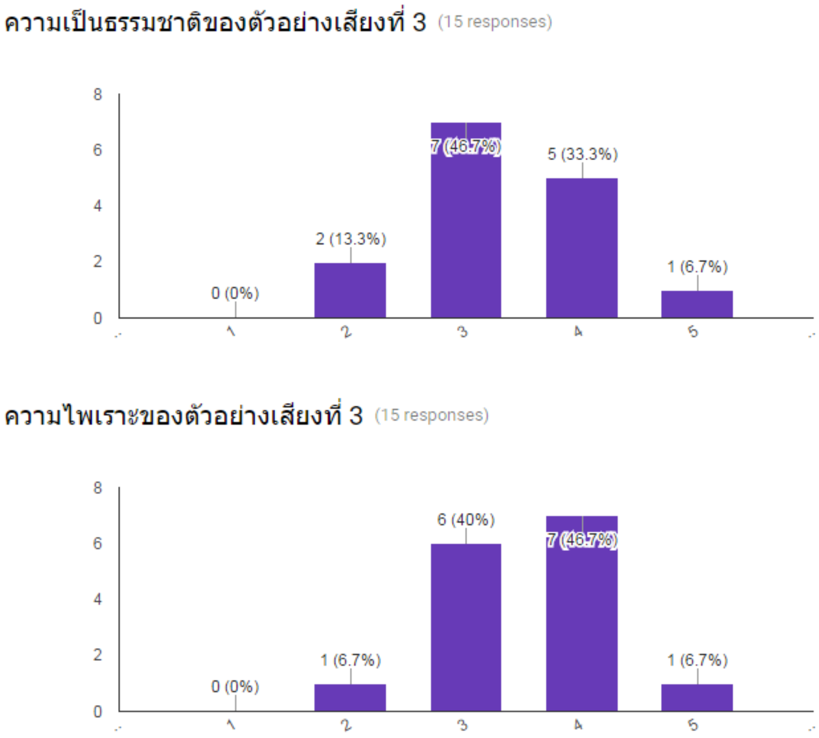


Figure 35 Result Graph of Sample 2 in Different Voice Duration under Same Phrase Context

- Sample 3

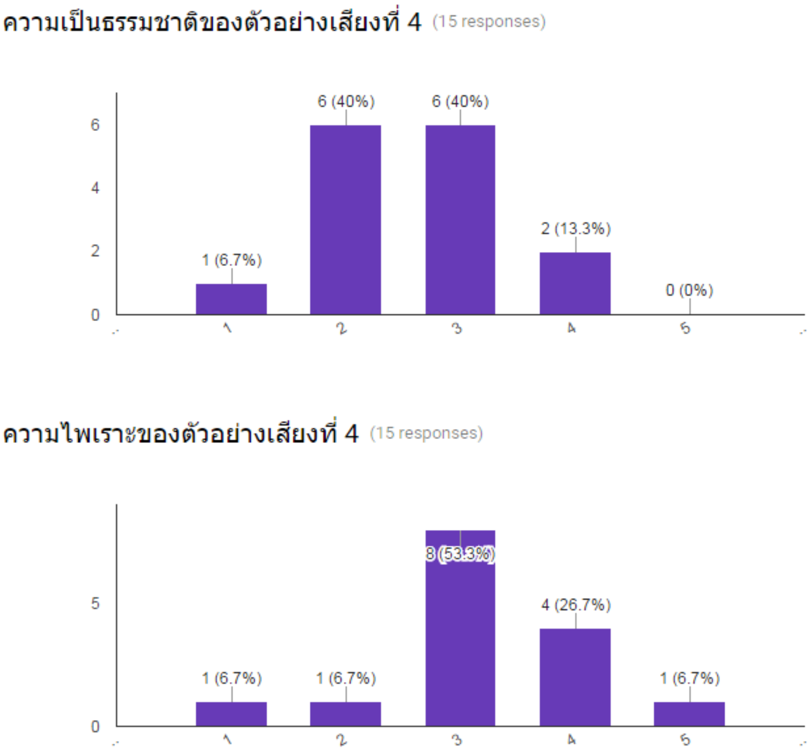


ความเป็นธรรมชาติของตัวอย่างเสียงที่ 3 (15 responses)

ความไพเราะของตัวอย่างเสียงที่ 3 (15 responses)

Figure 36 Result Graph of Sample 3 in Different Voice Duration under Same Phrase Context

- Sample 4



ความเป็นธรรมชาติของตัวอย่างเสียงที่ 4 (15 responses)

ความไพเราะของตัวอย่างเสียงที่ 4 (15 responses)

Figure 37 Result Graph of Sample 4 in Different Voice Duration under Same Phrase Context