UNIVERSITY OF SCIENCE AND TECHNOLOGY OF HANOI

UNDERGRADUATE SCHOOL

Research and development

BACHELOR THESIS

By

Nguyễn Đăng Hoà

Information and Communication Technology

Title:

HMM Training in Vietnamese Text-to-speech System

Supervisors: Prof. Dr. Vũ Tất Thắng

IOIT

INDEX OF CONTENTS

ACKNOWLEDGEMENTS

“I would like to express my gratitude to all those who gave me possibility to complete this thesis in specific as well as academic supporting in general that I received during the time of my studying at the University of Science and Technology of Hanoi.

Foremost, I would like to send my honest thanks to my supervisor Dr. Vũ Tất Thắng for his enthusiastic support and inspiration he gave me and other project’s participants during the internship. His advice helped us considerably in building certain knowledge about speech synthesise field. I am also really graceful to Dr. Lương Chi Mai for her guidance at the beginning of the internship.

I would like to thank Mr. Nguyễn Đức Thắng and Mr. Nguyễn Minh Tuấn for their collaboration in researching and building data during the project.”

LIST OF ABBREVIATIONS

TTS	Text-to-speech
HMM	Hidden Markov Model
X-SAMPA	Extended Speech Assessment Methods Phonetic Alphabet
IPA	International Phonetic Alphabet
ASCII	American Standard Code for Information Interchange
PDF	Probability Density Function
F0	Fundamental Frequency
ToBI	Tones and Break Indices

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

In this internship, we present a new approach to build a high-quality Vietnamese Text-To-Speech system. This approach is based on modelling our speech data by Hidden Markov Models, in order to apply the machine learning method to generate the artificial voice by concatenating piece of pre-recorded human speeches.

The data of Vietnamese speeches , however, takes much time to be built and cleaned, this thesis does not discuss the output of a Vietnamese TTS system, but the design system only. The results to be discussed is derived from the TTS system we built for general English, which demonstrates a reliable intelligibility with only 3% word error rate, but a lack of naturalness.

The project is still in process while 2000 sentences of Vietnamese speeches are being recorded and cleaned.

Key words: **Text-To-Speech, Vietnamese language, HMM, HTS, MaryTTS, Machine Learning**

I/ INTRODUCTION

1. Application of Speech Synthesis

Over the last few decades, speech synthesis (or TTS) has been greatly developed for both research and commercial application. While the researchers are making effort to implement the system that produces the most human-liked voices with the emergence of advances technology, some technology giants invest to integrate it into mobile devices. This fact renders the TTS system more common in various languages and applications, and remarkably better in the quality.

1.1 Applications for the blind

It appears that the most important real-life application of a TTS system is to support the blind to read and communicate with the electronic devices. Before synthetic speech, audio book is the only choice for the blind, but it is clear that having people to record any large book, such as a long novel, costs a lot of time and money.

Moreover, with the development of internet, reading books is no longer the only demand of blind users, but also interacting with the computer and other people in network. Speech synthesis brings the most important and longest application to the blind, which is screen-reading. *Thunder ScreenReader (Windows), Virtual Vision (Windows)* and *VoiceOver (Mac OSX)* are some typical screen-readers which are widely used by both blind and common users nowadays. And along with the development of speech recognition, people with reading impairment now are able to optimise their communication ability through internet without reading a single text.

1.2 Applications for Telecommunications and Multimedia

TTS system has been used for decades in all kinds of telephone enquiry systems. However, its quality just reached the acceptable level for a common customer in some

recent years when the generated voices get closer to human voice in both pronunciation and rhythm.

Another application of TTS is to help the drivers in reading emails and text messages. However, due to the complication of the TTS mobile-app, the number of languages supported is still limited.

2. Project Description

The project, lasts in 3 months, aims to build a high-quality Vietnamese TTS system to assist blind users in Vietnam to access written text on the computer. In order to work on this problem, some certain synthesiser technologies were investigated.

2.1 Synthesiser technologies

a. **Concatenation synthesis**

   Concatenation synthesis is based on concatenation of short samples (called units) of recorded speech. The duration of the units varies from 10 milliseconds up to 1 second, depends on the demanded accuracy of the system. Generally, concatenation synthesis produces the most natural-sounding speech, in comparison to human voice. Segmentation of the waveform, however, may result in some certain audible glitches in the output speech.

b. **Formant synthesis**

   Different to other types of speech synthesis, formant synthesis does not use human speech during the runtime. Instead, it computes some important parameters, such as fundamental frequency, voicing and noise levels, which then will be used to generate an artificial speech. The output of formant synthesis is normally far from human voice and sounds robotic. However, human-liked is sometimes not the first priority of a TTS

system when formant synthesis is able to produce a reliably intelligible speech, avoiding audible glitches at even the high speed.

c. HMM-based Synthesis

HMM-based synthesis is a method based on hidden Markov model - a statistical machine learning method. The system uses HMMs to model three main parameters of speech, which are frequency spectrum (vocal tract), fundamental frequency (voice source) and duration (prosody). The speech waveforms are then generated using maximum likelihood estimator.

In this internship, the HMM-based synthesis method was chosen to be researched and implemented for Northern Vietnamese language due to the naturalness and intelligibility it provides in the TTS system. In addition, because it is a machine learning method, the method saves us a lot of time, money and human resource.

II/ MATERIALS AND METHODS

1. Northern Vietnamese phonetics and phonology

Studying Vietnamese language is the first step in the project. Similar to normal human communication, one TTS system needs to understand the basic structure of one language in order to process and give the proper output speech. In this project, phonetics and phonology are two main characteristics that were investigated, since they are two main factors composing a speech.

1.1 Vietnamese syllable structure

In general, there has been different publications about Vietnamese syllable structure, written by both Vietnamese and international researchers. Most of them agree on the hierarchy structure of Vietnamese syllables (Thompson, 1987) (Doan, 1977) (Vogel et al., 2004), which describes Vietnamese "only single consonants in onset and coda positions, and a single vowel or a diphthong in the nucleus" (Vogel et al., 2004).

The publications of Thompson (1987) and Vogel et al. (2004) basically separate a Vietnamese syllable into two big components: onset and rhyme; and two smaller components, nucleus and coda, which both compose the rhyme component. This structure explanation is demonstrated in figure 1.1.

For example, the word "học" can be separated into three parts: the onset /h/, the nucleus /ɔ-6b/ and the coda /k/.

<table>
<tr><td rowspan="2">Onset
(Consonant)</td><td colspan="2">Rhyme</td></tr>
<tr><td>Nucleus
(Vowel)</td><td>Coda
(Consonant)</td></tr>
</table>

Figure 1.1. Basic structure of Vietnamese syllables (Thompson, 1987) (Vogel et al., 2004)

However, these assumptions fail to explain two important factors in Vietnamese phonetics structure, that are the existence of medial part and the effect of lexical tone on other parts rather than just the nucleus.

The hierarchy structure of Vietnamese syllables published by Doan (1977) solved the problem by adding another smaller part, called medial, into the rhyme, and covering the whole syllable with a lexical tone (as described in figure 1.2).

<table>
<tr><td colspan="4">Tone</td></tr>
<tr><td rowspan="2">Initial
(Consonant)</td><td colspan="3">Rhyme</td></tr>
<tr><td>Medial
/w/</td><td>Nucleus
(Vowel)</td><td>Ending
(Consonant)</td></tr>
</table>

Figure 1.2. Hierarchy structure of Vietnamese syllables (Doan, 1977)

Medial part of the syllables does not exist clearly in the normal written text, but does play a role in case that there appears a glide between an onset (initial) and a nucleus. The word “ngoan” [ŋwan] is a typical example where there appears the medial /w/, instead of the nucleus /ɔ/, which is clearly not pronounced in the word. Generally, the medial can be pronounced when the words contain any “uy”, “oa” or “qu” phones.

Lexical tone is another important part of Vietnamese syllable, which was not mentioned in the publications of Thompson (1987) and Vogel et al. (2004). In Vietnamese, lexical tone is indispensable factor that can even alter the meaning of the whole word. For instance, “ba” (father) - level tone and “bà” (grandmother) - falling tone are two different syllables with different meanings.

In Doan’s opinion, in Vietnamese, the lexical tone covers the whole syllable, from the initial to the ending part, which means the tone can affect the sound of all parts in the syllable. For example, the initial sound /s/ should be pronounced differently in two words “sau” (after) and “sáu” (six).

However, in 2005, Tran and his team carried a study using Diagnosis Rhyme Test (DRT) method to understand the effect of tone on each part of the syllable. The result, thus, rejects the information of tone on the initial consonant and the role of initial part in construction of syllable tone. Therefore, the final hierarchy structure of Vietnamese syllables was reached and used during the whole project, as described in the figure 1.3.

<table>
<tr><td rowspan="3">Initial
(Consonant)</td><td colspan="3">Rhyme</td></tr>
<tr><td>Medial
/w/</td><td>Nucleus
(Vowel)</td><td>Ending
(Consonant)</td></tr>
<tr><td colspan="3">Tone</td></tr>
</table>

Figure 1.3. Concluded hierarchy structure of Vietnamese syllables

1.2 Vietnamese phonology system.

Vietnamese phonology system varies from region to region, which means the system may not be the same at two different cities. In the project, Hanoi Vietnamese system is the only phonology system to be considered and used, since all the partakers are from Hanoi.

1.2.1 Initial consonants

Table 1.1 presents the categorisation of 19 initial consonants in Vietnamese phonology, proposed by Kirby (2011). According to Kirby, Hanoi Vietnamese initial consonants have some noticeable characteristics in comparison to other regions, that describe the similarity in pronunciation of phonemes “tr-” and “ch-”; “s-” and “x-”; as well as “d-”, “gi-” and “r-”.

	Labial	Labio-dental	Dental	Alveolar	Palatal	Velar	Glottal
Plosive	ɓ		t t^h	ɗ	tɕ	k	
Nasal	m		n		ɲ	ŋ	
Fricative		f v		s z		x ɣ	h
Approximant	w						
Lateral approximant			l				

Table 1.1. Hanoi Vietnamese initial consonants (Kirby, 2011)

1.2.2 Vowels and diphthongs

The nucleus is the main and indispensable part of a syllable. In Vietnamese, a nucleus is made up by a vowel or a diphthong - a combination of two vowels. In Kirby's (2011) and Doan's (1977) study, Hanoi Vietnamese distinguishes nine long vowels /i e ɛ a ɯ ɤ u o ɔ/, four short vowels /ɛ̆ ă ɤ̆ ɔ̆/ and three falling diphthongs /iə ɯə uə/.

	Front	**Central**	**Back**	
			Unrounded	Rounded
Close (High vowel)	i	iə ɯə uə	ɯ	u
Close-mid	e		ɤ ɤ̆	o
Open-mid	ɛ ɛ̆			ɔ ɔ̆
Open (Low vowel)	a ă			

Table 1.2. Hanoi Vietnamese vowels and diphthongs (Nguyen, 2016)

1.2.3 Lexical tones

Lexical tone is another important factor in Vietnamese language. A lexical tone can change the pitch of one syllable and the meaning of the whole word, for instance word "ba" (father) is different to word "bà" (grandmother).

In Vietnamese writing system, there are six lexical tones: (1) level tone, symbolised by the absent of any mark (e.g "ma"); (2) falling tone, symbolised by the mark (`) (e.g "mà"); (3) curve tone, symbolised by the mark (ʔ) (e.g "mả"); (4) broken tone, symbolised by the mark (~) (e.g "mã"); (5) rising tone, symbolised by the mark (´) (e.g "má"); (6) drop tone, symbolised by the mark (.) (e.g "mạ").

In this work, however, we distinguished Vietnamese into 8 tones. Aside from the first four tones: level, falling, curve and broken tone, the rising tone is separated into normal and sharp rising; the dropping tone is separated into normal and sharp dropping. The sharp

tones are symbolised by the same mark, but occur only in the syllables with “-c”, “-t”, “-p” and “-ch” endings and makes these syllables sound shorter.

Table 1.3 represents the full system and the notation of Hanoi Vietnamese lexical tones in the project.

Notation	Name		Register	F0 Conour	Duration
1	Ngang	Level	High-Mid	Level	Long
2	Huyền	Falling	Low	Slightly Falling	Long
3	Hỏi	Curve	Low	Falling	Long
4	Ngã	Broken	High	Falling-Rising	Long
5a	Sắc	Rising	High	Rising	Long
5b	Sắc	Rising	High	Sharply Rising	Short
6a	Nặng	Drop	Low	Dropping	Short
6b	Nặng	Drop	Low	Sharply Dropping	Short

Table 1.3. Hanoi Vietnamese lexical tones

1.2.4 X-SAMPA - Phonetics ASCII representation

In order to represent the phonetic symbols in the computer, a world-wide standard called X-SAMPA was set. This standard contains the table to convert phonetic symbols into ASCII form as shown in the table 1.4a and 1.4b.

Vowel type	Grapheme (Orthography)	Phoneme		Vowel type	Grapheme (Orthography)	Phoneme	
		IPA	X-SAMPA			IPA	X-SAMPA
Long vowel	a	a	a	Short vowel	ă, a (au, ay)	ă	a_X
	e	ɛ	E		â	ɤ̆	7_X
	ê	e	e		a (anh, ach)	ɛ̆	E_X
	i, y	i	i		o (ong, oc)	ɔ̆	O_X
	o, oo	ɔ	O	Diphthong	ia, iê, yê, ya	iə	i@
	ô, ôô	o	o		ua, uô	uə	u@
	ơ	ɤ	7		ưa,ươ	ɯə	M@
	u	u	u				
	ư	ɯ	M				

Table 1.4a. Vietnamese vowels X-SAMPA table

No.	Graph -eme	Position	Phoneme		No.	Graph -eme	Position	Phoneme	
			IPA	X-SAMPA				IPA	X-SAMPA
1	b	initial only	ɓ	b	14	t	initial, final	t	t
2	đ	initial only	ɗ	d	15	p	initial, final	p	p
3	x , s	initial only	s	s	16	n	initial, final	n	n
4	g, gh	initial only	ɣ	G	17	ch	final after i,ê,a	k͡+	k_+
5	kh	initial only	x	x	18	c	final after u,o,ô	k͡p	kp
6	l	initial only	l	l	19	ch,c	final except 17, 18	k	k
7	v	initial only	v	v	20	m	initial, final	m	m
8	th	initial only	t^h	t_h	21	nh	final after i,ê,a	ɲ͡+	N_+
9	d,gi,r	initial only	z	z	22	nh	initial only	ɲ	J
10	ph	initial only	f	f	23	ng,ngh	initial	ŋ	N
11	tr, ch	initial only	tɕ	ts\	24	ng	final after u,o,ô	ŋ͡m	Nm
12	h	initial only	h	h	25	ng	final except 24	ŋ	N
13	c,k,q	initial only	k	k					

Table 1.4b. Vietnamese consonants X-SAMPA table

2. HMM-based Text-to-speech system.

2.1 Hidden Markov Model

Hidden Markov model is a statistical model used widely to represent the probability distributions over a sequence of observations. In a hidden Markov, the system is modelled as a Markov process with a finite sequence of hidden states, which are not directly visible; and dependent outputs, which are observable.

It is assumed that each hidden state gives a certain probability distribution over the output tokens. As a results, by observing the sequence of discrete time ouputs, an HMM can give information about the sequence of hidden states, by which we can compute the parameters of the model.

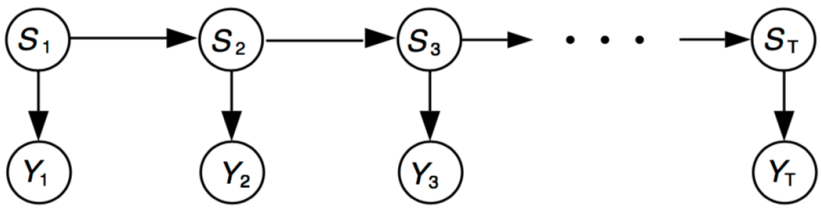

Figure 2.1. Example of an HMM - *S is the hidden state, Y is the observation* (Ghahramani, 2001)

Figure 2.1 demonstrates a T-state left-to-right model, in which $\{S_1, S_2, \ldots S_T\}$ is the set of hidden states and $\{Y_1, Y_2, \ldots Y_T\}$ is the set of output observations. An HMM must satisfy three properties. First, the output Y_t depends only on the state S_t. Second, given the value of S_{t-1}, the current state S_t is independent to all other states except t-1. In other words, the state at any given time reveals us all about the history of the process, so that the future can be predicted without needing to know the value of previous states. Finally, the value set of hidden state S_t is discrete.

The model parameter of the first-order HMM can be factored in the following way:

$$P(S_{1:T}, Y_{1:T}) = P(S_1)P(Y_1|S_1)\prod_{t=2}^{T} P(S_t|S_{t-1})P(Y_t|S_t),$$

where the notation $X_{1:T}$ means $X_1,\ldots, X_T$; $P(S_1)$ is the initial state distribution; $P(S_t|S_{t-1})$ is the transition distribution and $P(Y_t|S_t)$.

The output probability distribution $P(Y_t)$ of the the observational data Y at state t can be either discrete or continuous depending on the observations, and thus, can be modelled by different Probability Density Function (PDF). For instance, in the continuous distribution HMM with continuous observation data, the output of each state is usually modelled by a mixture of multivariate Gaussian distribution (Figure 2.2).

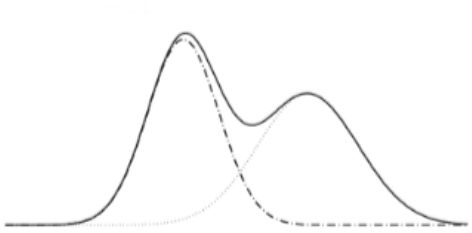

Figure 2.2. Multi Mixture Gaussian PDF

Hidden Markov Model in speech synthesis

In speech processing, speech parameter sequence is often modelled as a left-to-right HMM. Spectral, f0 and duration are three parameters to be modelled in an HMM-based speech synthesis system.

Spectral, which is a continuous variable, can be modelled by continuous HMM in the same way as speech recognition system. F0, however, is continuous in the "voiced" region and discrete in "unvoiced" region, which renders it much more complicated to use either the discrete and continuous HMMs to model. Yoshimura (2002) proposed another type of HMM in order to solve the problem, that is multi-spaced HMMs. Figure 2.3 demonstrates the structure of HMM for spectral and f0 parameters.

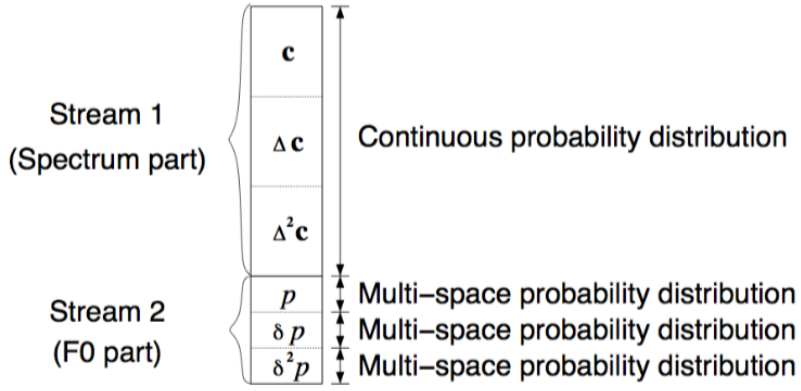

Figure 2.3. Structure of feature vectors modelled by HMMs (Yoshimura, 2002)

State duration densities in HMM-based speech synthesis, on the other hand, is impossible to be modelled by single Gaussian distributions estimated from histograms of the state duration, since the variance of distribution for phonemes which appear only once in the training data is unobtainable. In the work of Yoshimura et al. (2002), state duration of each phoneme HMM are regarded as a multi-dimensional observation in order to overcome the problem. In other words, state durations of each phonemes HMM is

modelled by multi-dimensional Gaussian distribution. Number of dimension in state

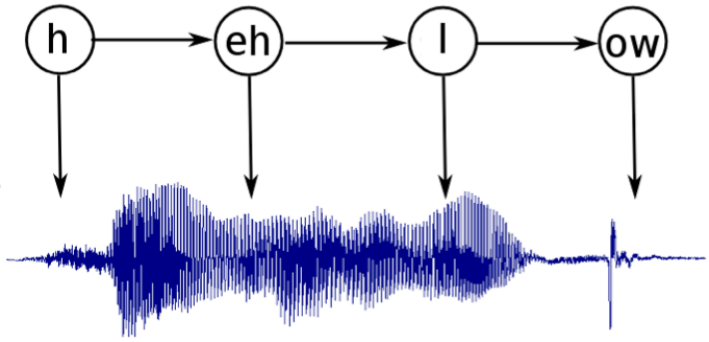

Figure 2.4. HMM of a speech synthesis system.

duration densities is the number of state of HMMs.

Figure 2.4 explains simply a model of a synthesis system, where the sequence of phonemes is the hidden states and the waveform represents the output observations. In an HMM-based speech synthesis system, a phoneme and its dependent contextual features must be represented by an HMM, e.g the phoneme */th/* in the sentence "*The cat is on the mat.*" should be represented as *<phone=/th/, next_phone=/ax/, word='the', next_word='cat', num_syllables=6…>*.

The representation above can be regarded as the context dependent duration models, and is used, along with the obtained coefficients and a decision tree, to train the context dependent HMMs. Figure 2.5 represents the basic workflow of an HMM-based speech synthesis system.

In the synthesis part, an arbitrary input sequence of text is synthesised to be converted to context-based label sequence, which then is used to create a sentence HMM. Subsequently, state durations of the obtained HMM are estimated using maximum likelihood to the state duration densities. Finally, speech is synthesised from the generated mel-ceptrum and F0 parameter using Mel Log Spectrum Approximation filter.

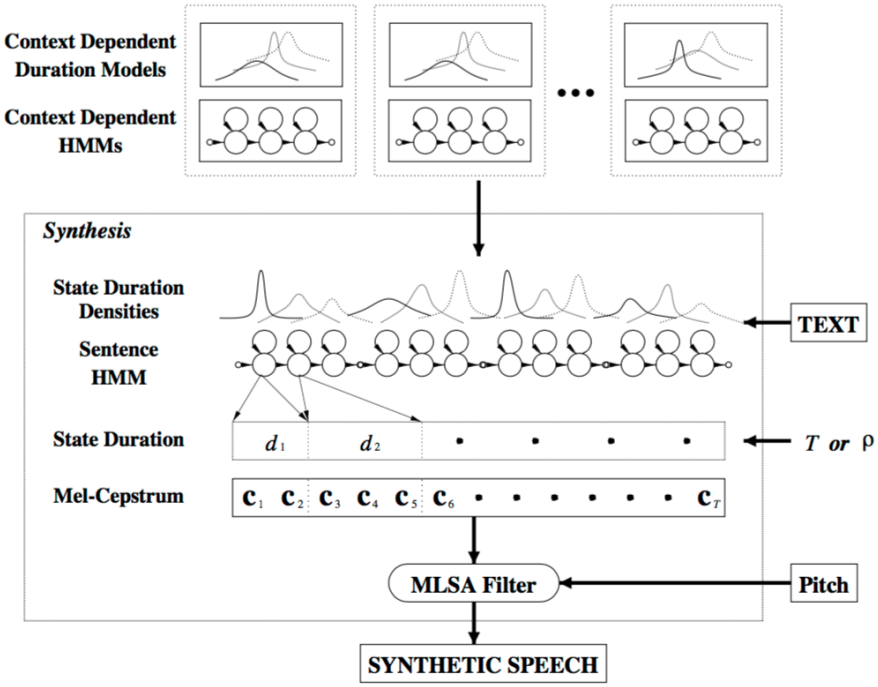

Figure 2.5. Workflow of speech synthesis system.

2.2 Contextual features

In HMM-based speech synthesis, contextual features considerably affect the spectrum, F0 and duration of a phoneme. In the HTS, a toolkit for HMM-base speech synthesis developed by Tokuda et al. (2002), the contextual features being considered were:

a. Phoneme

- current phoneme
- preceding and succeeding two phonemes
- position of current phoneme in the current syllable

b. Syllable

- numbers of phonemes in preceding, current and succeeding syllables
- stress and accent of preceding, current and succeeding syllables
- position of current syllable in current word and phrase

- numbers of preceding and succeeding stressed syllables within current phrase
- numbers of preceding and succeeding accented syllables within current phrase
- number of syllables from previous stressed syllable
- number of syllables to next stressed syllable
- number of syllables from previous accented syllable
- number of syllables to next accented syllable
- vowel identity in current syllable

c. Word

- guess at part of speech of preceding, current, and succeeding words
- numbers of syllables in preceding, current, and succeeding words
- position of current word in current phrase
- numbers of preceding and succeeding content words in current phrase
- number of words from previous content word
- number of words to next content word

d. Phrase

- numbers of syllable in preceding, current and succeeding phrases
- position of current phrase in major phrases
- ToBI endtone of current phrase syllable

e. Utterance

- numbers of syllable, words, and phrases in utterance

In Vietnamese language, the contextual features about the stress of the syllable within a word can be skipped because there is no boundary between a syllable and a grapheme, and all syllables can be stressed the same in a natural Hanoi Vietnamese speech.

3. Architecture of TTS

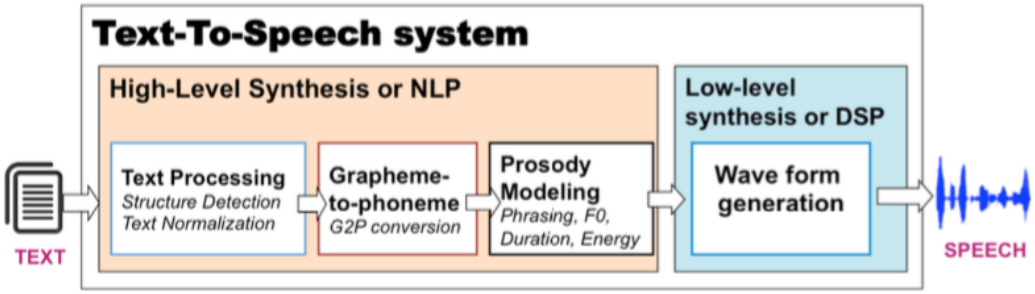

Figure 3.1. Basic architecture of text-to-speech system (Nguyen, 2016)

In Nguyen's work (2016), a basic architecture of a text-to-speech system has two main parts and four components, as described in the figure 3.1. The former part - Natural Language Processing (NLP) is made up of three components: text processing, grapheme-to-phoneme and prosody modelling. The latter part implements the Digital Signal Processing (DSP) to generate the speech waveform.

3.1 Text processing

Text processing is divided into two steps: structure detection and text normalisation.

In the former step, the input text is segmented into sentences, phrases and words with the proper attached tags. The segmentation can be done by determining the white spaces, punctuations and other delimiters of the input text.

The text normalisation step is responsible for converting the Non-Standard-Words (NSWs) into the speakable forms. Since the real input text may contain some NSWs, such as numbers, date, currency or acronyms, a TTS system must be able to speak it in the

correct form of Vietnamese language. For example, a date written in the form *"DD-MM-YYYY"* should be converted into a natural speakable form as *"Ngày DD Tháng MM Năm YYYY"*.

3.2 Grapheme-to-phoneme conversion

G2P conversion is the second block in the architecture of a standard TTS system. In this work, a Vietnamese tonophone system was used during the conversion step. The system was represented as a xml file *allophones.vi.xml,* which defines 19 initial consonants and 170 rhymes (combination of vowels and lexical tones).

This step takes the processed text from the previous block (text processing) and produces the sequence of phonemes as the output by using Vietnamese letter-to-sound rule or by looking up a dictionary. This sequence will be then synthesised based on the trained HMMs in order to generate the expected waveform.

3.3 Prosody Modelling

At this point, the speech parameters, including spectral, F0 and duration will be predicted based on the trained Hidden Markov Model. The sequences of phonemes obtained from the second block can be regarded as the sequence of hidden states; the speech parameters, on the other hand, are predicted using Maximum Likelihood algorithm.

3.4 Speech Synthesis

Speech synthesis uses low-level digital signal processing to generate the speech waveform from the parameters obtained from previous block. In general, there are two common approaches in this step: (1) source/filter synthesiser: produce an artificial piece of speech using source/filter model from the parametric representation of speech; (2) Concatenation synthesiser: concatenate the pre-recorded real human voice to construct the speech.

In this work, we use the concatenation synthesiser approach, as it produces higher quality speech and can take advantage of the Machine Learning method using HMMs.

4. Method

4.1 HTS Framework

In this work, we use the HMM-based Speech Synthesis System (HTS) framework developed by Tokuda et al. (ver 2.3, 2015). HTS is a modified version of HMM Toolkit (HTK), which was built by Machine Intelligence Laboratory of the Cambridge University Engineering Department. It consists of a set of library modules and tools available in C source form, which provide sophisticated facilities for speech analysis, HMM training, testing and results analysis.

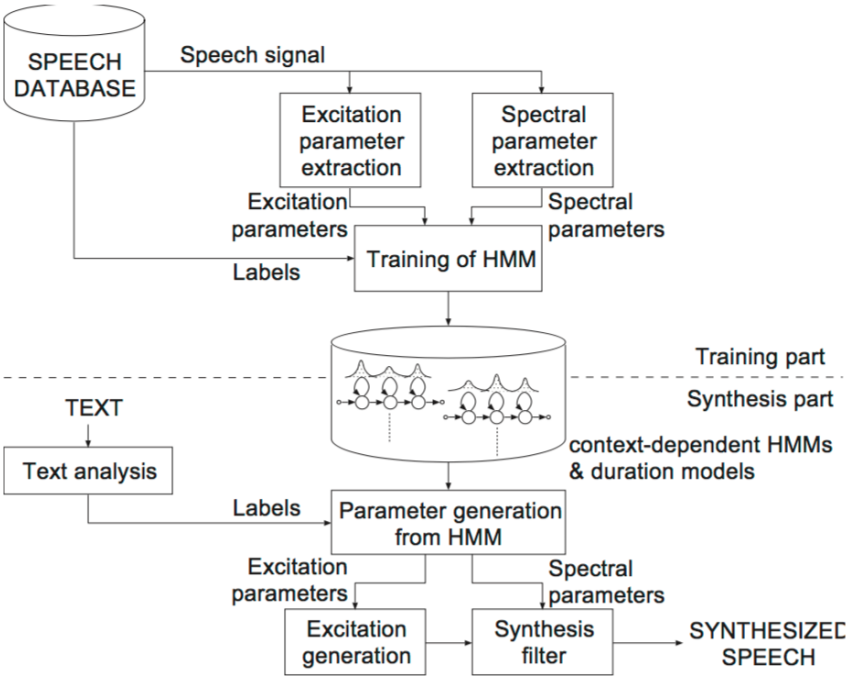

Figure 4.1. Basic Workflow of HTS

In general, HTS has two main parts: training part and synthesis part, as illustrated in figure 4.1. In the training part, the spectral and excitation (log F_0) parameters are

extracted from the speech database before being modelled by context-dependent HMMs. The synthesis part does the inverse process, which firstly takes an arbitrary text as an input. Second, the text is then converted into context-dependent label sequence, which is subsequently used to construct an utterance HMM. Third, the state duration, spectral and F0 parameters are generated that maximise their output probabilities. Finally, a speech waveform is generated based on the sequence of obtained parameters.

4.2 MaryTTS Framework

MaryTTS is an open-source, multi-lingual TTS platform written in Java, which was developed by German Researcher Centre for Artificial Intelligence (DFKI). The platform provides us some useful tools and Graphic User Interface (GUI) for all step in the HTS that makes us easier to implement. Plus, MaryTTS supports a quick tutorial for training and building a new language speech synthesis system, and now has a big development community on GitHub.

Figure 4.2 represents a basic workflow of MaryTTS system.

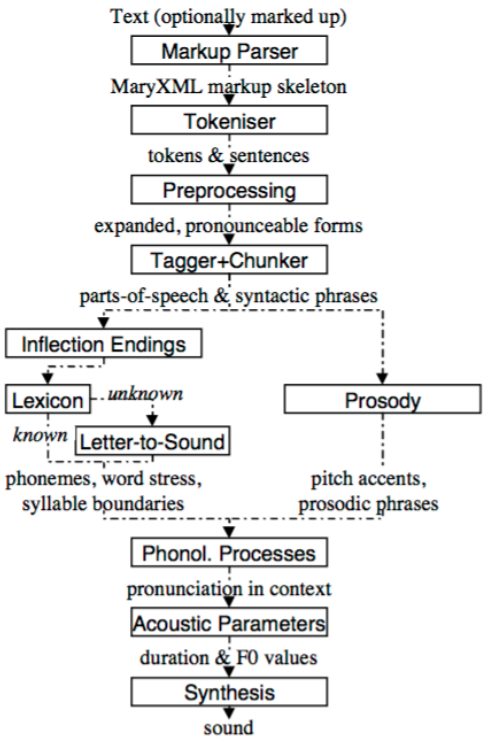

Figure 4.2. Basic Workflow of MaryTTS

MaryXML

As MaryTTS works on a huge resource of data of speech for training process, these data must be represented in a standard format. The MARY system uses an internal XML-based language for this representation, called MaryXML. The syntax of a MaryXML reflects the informations required in various processing step in the MaryTTS system.

Optional Markup Parser

The MaryTTS can take either a plain text or a markup language (e.g SABLE) as the input for the speech synthesis. However, most of the information to be represented in MaryXML, is too detailed to be expressed by tags defined in input markup languages, such as the location and types of stressed phonemes. Consequently, MaryTTS builds the Markup Parser to translate other XML formats into MaryXML format for the MaryTTS to understand during the training and synthesis process.

Tokeniser

Tokeniser divides the plain text into tokens, i.e words within a sentence. Each token is enclosed by a <*t*>...</*t*> MaryXML tag. All local information about a token is added inside the tag <*t*> as the attributes. In addition, the <*div*>...</*div*> MaryXML tag is used to enclose a sentence, which is determined by punctuation marks (e.g dot, question mark).

Preprocessing

Preprocessing module works as a text normalisation, which converts the number or abbreviation tokens into the pronounceable forms. The numbers are expanded to the written form, for example “1” is converted to “một”. The abbreviations are divided into two groups: those are spelled out (e.g “www” is converted into “vê kép vê kép vê kép”), those should be expanded (e.g “THCS” is converted into “Trung học cơ sở”).

Speech tagger and chunker

Speech tagger is performed to define the part-of-speech, while chunk parser determine the boundary of noun phrases, prepositional phrases and adjective phrases. The information about part-of-speech and chunk is added inside the tag <*t*> in the MaryXML file.

Phonemisation

The phonemisation works as a Grapheme-to-phoneme module in the MaryTTS platform. It requires a XML file that defines the allophone system of Vietnamese. As described above, this file contains 19 initial consonants and 170 rhymes of Vietnamese phonetic system. MaryTTS provides a GUI for the transcription process, that also enable us to automatically transcript a large number of words based on the training of existent data.

HMM Training for new language

MaryTTS provides Voice Import Tools, enables us to import the speech data in WAV format. 16-bit mono 16000 Hz WAV files are often preferred in the MaryTTS platform, but other properties are also possible. Voice Import Tool also takes the input of the transcription in both grapheme and phoneme of all speech data files.

Before using being able to use this tool, users must have some following dependent tools:

- Praat - For pitch marks
- Edinburg Speech Tools - For MFCCs and Wagon (CART)
- Festvox version 2.1 or newer. This tool provides the EHMM labeller, which is used for automatic labelling

In this project, we recored 2000 WAV files as the training data. All the recording is done with the built-in microphone in the Macbook Pro 2015 and in Hanoi Vietnamese accent.

III/ RESULTS AND DISCUSSION

Since the data of Vietnamese voice requires at least 10,000 word transcriptions and 1000 recorded sentences, it takes us much time to work on and is still being processed. Thus, in this chapter, the results and discussion will be derived from our first approach to MaryTTS for general English.

In this test, we use the ARCTIC data from CMU that includes 1132 WAV recorded sentences and a XML file of English allophones system. The result can be discussed based on two main criteria of a TTS system: intelligibility and naturalness. In order to test the results, we let 16 participants listen to 50 random speeches among 1132 generated files. These users assessed the output on the scale of 0-5. The result is then collected and compared to the outputs of diphone/formant synthesiser, which was retrieved from the website of Institute of Maschinelle Sprachverarbeitung and tested the same way.

All the survey-participants are used to listening to English.

1. Intelligibility

The intelligibility of a system can be evaluated by the word error rates. Each generated speech will be played only once, the survey-participants are asked to list the words with incorrect or unnatural pronunciation.

After the survey, the results are collected and described as below:

- Average number of error words per sentence: 0.625, similar to 3.5% error word rate.
- Average score of intelligibility: 3.875.

The results reflect a relatively reliable intelligibility of general English TTS system, while most of the words are pronounced correctly.

In addition, in comparison to the formant synthesiser, the output of the HMM-based TTS system was reported much clearer and more intelligible, since most of HMM-based output sentences were understood after one play, but the formant output.

2. Naturalness

The naturalness of a system is assessed by the human-liked level of the generated speeches. The survey-participants were allowed to evaluate the naturalness of the output sentences in four types: declarative sentences, imperative sentences, exclamatory sentences and interrogative sentences in the scale of 0 to 5.

The survey provides us the average score of naturalness at only 2.95. Although it was rated much more natural than the output of formant synthesiser, the score demonstrates that it is still a challenge to produce a human-liked TTS system despite the huge resource of training data.

This issue can be tackled by either training a larger amount of speech data or providing the system more various of contexture features.

IV/ CONCLUSION

The internship gave us a wealth of experience in research about speech processing, Vietnamese Natural Language and Machine Learning. Although we have not built a sufficient resource of Vietnamese speech data to be trained, the approach in building an HMM-based TTS system proved its potential and reliability with a test on general English voice data. The project, however, can be finished in the future when the data is built.

Aside from research skills, the internship provides us a good environment to train our teamwork ability, time management and task assignment. Three of us has worked in a team in three months, which helped us understand the importance of the collaboration in research.

Finally, working in the project of speech processing provides us certain skills in programming various languages. During the internship, we must be able to handle three main languages, which are Python, C and Java. This considerably enhance our ability to adapt training a new language when going to work in the future.

REFERENCES

Dinh Anh-Tuan, Phan Thanh-Son, Vu Tat-Thang and Luong Chi-Mai. Vietnamese HMM-based Speech Synthesis with prosody information. *8th ISCA Speech Synthesis Workshop*, Barcelona, Spain, August 2013.

Ghahramani, Z. An introduction to Hidden Markov Models and Bayesian Networks. *International Journal of Pattern Recognition and Artificial Intelligence*, London, England, 2001.

Kirby J., Illustrate of the IPA Vietnamese (Hanoi Vietnamese). *Journal of the International Phonetic Association,* 41(03):381–392, 2011

Pammi S., Charfuelan M., Schröder M.. Multilingual Voice Creation Toolkit for the MARY TTS Platform, Language Technology Lab Saarbrücken and Berlin, Germany, 2010.

Schröder M., Charfuelan M., Pammi S. and Steiner I.. Open source voice creation toolkit for the MARY TTS Platform, Language Technology Lab Saarbrücken and Berlin, Germany, 2011.

Thi Thu Trang Nguyen. HMM-based Vietnamese Text-To-Speech : Prosodic Phrasing Mod- eling, Corpus Design System Design, and Evaluation, Université Paris Sud - Paris XI, Jan 2016.

Yamagishi J. An introduction to HMM-Based Speech Synthesis, University of Edinburgh, Edinburgh, October 2006.

Yamagishi J., Zen H., Wu Y., Toda T. and Tokuda K..The HTS-2008 System: Yet Another Evaluation of the Speaker-Adaptive HMM-based Speech Synthesis System in The 2008 Blizzard Challenge, University of Edinburgh, Edinburgh, 2008.

Zen H., Nose T., Yamagishi J., Sako S., Masuko T., Black A. W. and Tokuda K.. The HMM-based Speech Synthesis System (HTS) Version 2.0. *6th ISCA Workshop on Speech Synthesis*, Bonn, Germany, August 22-24, 2007.