# DEVELOPMENT OF VIETNAMESE TEXT TO SPEECH SYSTEM USING MARYTTS

**Van-Dong Pham[1,2], Dang-Khoa Mac[1], Do-Dat Tran[3]**

[1] International Research Institute MICA, Hanoi University of Science and Technology
[2]Faculty of Information Technology, Hanoi University of Mining and Geology
[3] Ministry of Science and Technology, Vietnam
*phamvandong@humg.edu.vn, dang-khoa.mac@mica.edu.vn, tddat@most.gov.vn*

**Abstract**

Development of text-to-speech (TTS) systems for a language normally require a lot of work, such as building a specific speech database, doing many linguistic researches (i.e. phonetics, phonology, prosody) on the target language. For Vietnamese, building a TTS system is an important and difficult work. This paper presents the experiment of building a TTS system for Vietnamese language, a primary language of Vietnam, using MaryTTS framework. The characteristics of Vietnamese language was studied for using to build the TTS system, thank to many linguistic researches. The Vietnamese synthesized speech was evaluated by Vietnamese speakers in perception tests. The result shows that the synthesized speech have a good quality, compare with other Vietnamese TTS systems.

**Keywords:** Text-to-speech, statistical approach, Vietnamese language.

## 1 Introduction

Speech synthesis is the artificial production of human speech. A computer system used for this purpose is called a speech computer or speech synthesizer, and can be implemented in software or hardware products. A text-to-speech (TTS) system converts normal language text into speech; other systems render symbolic linguistic representations like phonetic transcriptions into speech [1].

The two primary technologies generating synthetic speech waveforms are concatenative synthesis [2], [3] and formant synthesis [4]. Each technology has strengths and weaknesses, and the intended uses of a synthesis system will typically determine which approach is used.

HMM-based synthesis [5], [6] is a synthesis method based on hidden Markov models, also called Statistical Parametric Synthesis. In this system, the frequency spectrum (vocal tract), fundamental frequency (voice source), and duration (prosody) of speech are modeled simultaneously by HMMs. Speech waveforms are generated from HMMs themselves based on the maximum likelihood criterion.

Deep learning-based synthesizers use Deep Neural Networks (DNN) [7]–[9], which are trained on recorded speech data. Some DNN-based speech synthesizers are approaching the quality of the human voice. Examples are WaveNet by DeepMind, Tacotron by Google and Deep Voice (which uses the WaveNet technology) from Baidu.

For Vietnamese, some voice synthesis systems have also been introduced in recent years: MICA TTS [10]–[12], VOS, vnSpeak TTS, eSpeak … They use the concatenative and HMM technology. Recently, Viettel [13] and FPT [14] are focus on the Deep Neural Networks to build Vietnamese TTS.

There are many Vietnamese TTS systems but voice quality is not as good as other voices. We use statistical approach, a good and polular method to build Vietnamese TTS using MaryTTS framework. MaryTTS is an open-source, multilingual Text-to-Speech Synthesis platform written in Java. It was originally developed as a collaborative project of DFKI's Language Technology Lab and the Institute of Phonetics at Saarland University. As of version 5.2, MaryTTS supports German, British and American English, French, Italian, Luxembourgish, Russian, Swedish, Telugu, and Turkish; more languages are in preparation.

We use MaryTTS becourse it comes with toolkits for quickly adding support for new languages and for building unit selection and HMM-based synthesis voices.

A very general architecture of a TTS is show in figure 1. The Text Processing component handles the transformation of the input text to the appropriate form so that it becomes

speakable. The G2P Conversion component converts orthographic lexical symbols into the corresponding phonetic sequence. The Prosody Modeling attaches appropriate pitch, duration and other prosodic parameters to the phonetic sequence. Finally, the Speech Synthesis component takes the parameters from the fully tagged phonetic sequence to generate the corresponding speech waveform [15].

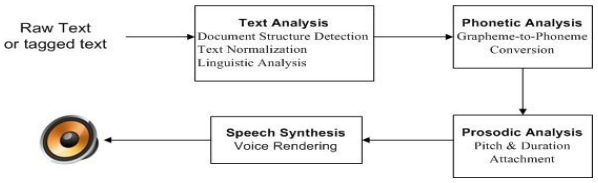


Figure 1 *Basic system architecture of a TTS system* [16]

In this paper we will describe our work of building a pilot Vietnamese TTS using MaryTTS framework. To build Vietnamese we study on phonetics and Vietnamese grammar, it is presented in Section 2. In section 3, the deployment of Vietnamese TTS using MaryTTS is described. The evaluation of the system performance with the proposal method is evaluated in the fourth section. The paper ends with some discussions and conclusions.

## 2 Vietnamese language

Vietnamese [17], is the language of the Vietnamese (Kinh people) and is the official language in Vietnam. This is the mother tongue of about 85% of the Vietnamese population, along with more than four million Vietnamese oversea. Vietnamese is also the second language of ethnic minorities in Vietnam. Although Vietnamese has a number of borrowed words from the Chinese language and previously used Nôm (a Chinese-based script) to write, Vietnamese is considered one of the languages of the Austroasiatic faminly that has the number of people who speak the most.

In Vietnamese, phonetic units and grammatical units (syllables and morphs) are the same. In addition, each syllable in Vietnamese has a stable and complete structure consisting of distinct sound units such as monosyllabic. Therefore, the role of the syllables in Vietnamese is much different than that of European languages.

According to Đoàn [18] and many other authors, Vietnamese syllables are a unit with a hierarchical structure. This hierarchical structure is represented by the following syllabus structure:

<table>
<tr><td rowspan="2"></td><td colspan="3">Tone</td><td></td></tr>
<tr><td rowspan="2">Initial</td><td colspan="3">Rhyme</td><td></td></tr>
<tr><td>Medial</td><td>Nucleus</td><td>Coda</td></tr>
</table>

*Figure 2* ***Vietnamese syllabus structure*** [18]

The first Vietnamese phonetics document was produced by [19] with modern spelling analysis. Other important descriptions including [20]–[23]. Then Vietnamese phonetics has been the subject of much attention and debate by many other researchers such as [18], [24]–[30].

According to Kirby [30], Hanoi Vietnamese has the 19 initial consonants, combined with the study of [18] we can describe the initial consonant in the following figure.

| | | | Labial | Labio-dental | Dental | Alveolar | Palatal | Velar | Glottal |
|---|---|---|---|---|---|---|---|---|---|
| Plosive | Unaspirated | Voiced | ɓ (**ba**) | | | ɗ (**đa**) | | | |
| | | Voiceless | | | t (**ta**) | | | k (**ca**) | ʔ (**à**) |
| | Aspirated | | | | tʰ (**tha**) | | tɕ (**cha**) | | |
| Nasal | | | m (**ma**) | | n (**na**) | | ɲ (**nhà**) | ŋ (**nga**) | |
| Fricative | Voiced | | | v (**va**) | z (**da**) | | | ɣ (**gà**) | |
| | Voiceless | | | f (**pha**) | s (**xa**) | | | x (**khá**) | h (**hoa**) |
| Approximant | | | w (**oan**) | | | | | | |
| Lateral approximant | | | | | l (**là**) | | | | |

Figure 3 Hanoi Vietnamese inital consonants

The phonological contrast between the idiopathic/non-idiomatic features of the syllables has created distinctive features that make up the actual content of two phonemes : a phoneme is a vowel /-u̯-/, opposite of another negative term phoneme /zero/. Both /- u̯-/ (may be written as /-w-/) and /zero/ play the role of the medial sound.

Medial sound only functions to attenuate the tone of the syllable after the opening rather than creating the main timbre of the syllable, so that a positive active phonetic accompaniment to this component can only be a glide, a semi-non-verbal chirps. This means that the sound is unlikely to be at the initial of the syllable to combine with the tone of the syllable [18].

Arcoding to Kirby [30] Ha Noi Vietnamese has 12 final syllables according to the image below.

| | Bi-labial | Labial-dental | Dental | Alveolar | Labial-velar | Pre-velar | Velar |
|---|---|---|---|---|---|---|---|
| Plosive | p | | | t | k͡p (final after o,u,ô) | k̟ (final after i,ê,a) | k |
| Nasal | m | | | n | ŋ͡m (final after i,ê,a) | ŋ̟ (final after i,ê,a) | ŋ |
| Approximant | | w | j | | | | |

Figure 4 Hanoi Vietnamese final consonant

In the Vietnamese syllables, the « điểm thanh tính » is always vowel, so the tones are always struck on the top of the vowel as the main sound. Consonants can not be made into syllables and syllables never occur in consonant clichés but only in vowel tones. The vowel at the top of the syllable always brings the dominant melody of the syllable. Unless the timbre is depreciated by the medial [-w-] or ends with a semi-vowel, the timbre consonant is pronounced from the beginning to the end of the syllable. Because of these reasons, the vowel in the Vietnamese syllabus is considered the main syllable or the nucleus of the syllable. In other words, without vowels, no syllables.

Vietnamese Hanoi distinguishes eight tones, six syllables in open syllables or have the last resonance sound, and a two tone pattern in the syllables that stop the mouth from pronouncing [27].

| | | | | | |
|---|---|---|---|---|---|
| *ngang* | A1 | ˧ (level) | ma˧ | *ma* | 'ghost' |
| *huyền* | A2 | ˨ (mid falling) | ma˨ | *mà* | 'but, yet' |
| *sắc* | B1 | ˧˥ (rising) | ma˧˥ | *má* | 'cheek' |
| | D1 | ˧ (rising checked) | mat˧ | *mát* | 'cool' |
| *nặng* | B2 | ˨˩ (low glottalized) | ma˨˩ | *mạ* | 'rice seedling' |
| | D2 | ˩ (low checked) | mat˩ | *mạt* | 'louse, bug' |
| *hỏi* | C1 | ˩˧ (low falling) | ma˩˧ | *mả* | 'tomb' |
| *ngã* | C2 | ˧˩ (broken) | ma˧˩ | *mã* | 'code' |

Figure 5 Tone of Hanoi Vietnamese

The following section details the construction of Vietnamese TTS using MaryTTS.

## 3 Vietnamese TTS development using Mary TTS framework

This section presents the deployment of the Vietnamese TTS system using MaryTTS framework.

### 3.1 Mary TTS

The earliest version of MaryTTS was developed around 2000 by Marc Schröder as a collaborative project of DFKI's Language Technology Lab and the Institute of Phonetics at Saarland University. For many years, it evolved and matured, first as an in-house Text-to-Speech (TTS) component, and subsequently as a fully open-source TTS platform with a growing community. Since its origins, MaryTTS has changed significantly, but it remains true to its original goal: a ***M**odular **A**rchitecture for **R**esearch in s**Y**nthesis* written in Java.

The properties of the MARY system are explained here along two lines: on the one hand, the architecture of the system from a natural language processing point of view; on the other hand, the workings of the system from a technical viewpoint.

### 3.2 Corpus

We built the Vietnamese TTS system using a database of 2316 Vietnamese words recorded by Mica with the voice of Le Diem. Here are some examples:

( TTSCorpusF01_lediem_GEN_0001 "chức trách của lê duật , là giữ đất đai , thấy giặc đến không thể không báo cáo , nhưng y cũng là một tay tướng giỏi" )

( TTSCorpusF01_lediem_GEN_0002 "nhìn xóm chài lúc sẩm tối càng thêm hiu hắt và buồn , các nóc thuyền đều rục rịch" )

( TTSCorpusF01_lediem_GEN_0003 "đến đêm , xóm chài cùng quây quần thành xóm nổi , ánh đèn dầu le lói , yếu ớt , khiến tôi có cảm giác , đời họ cũng nhàng nhàng và tôi tối" )

Audio recorded in wave file format, txt.done.data file records the text data of sentences.

### 3.3 Development

We divided the system into two parts, including natural language (left side) and speech processing (the right side). With the input text, combined with the set of allophones provided from the beginning (allophones.vi.xml), this text will be separated into sentences, then split into tokens. Each language is separated, for example English, the token is an integral word, and other forms are converted to standard form, such as playing -> play. Once separated into separate tokens, the system will standardize the text, for example: "57" -> "năm mươi bảy", "HCM"->"Hồ Chí Minh", "20/10" -> "Hai mươi tháng mười" … Once the text has been normalized, the words of the text will be converted into the phones, for example: "Trũi khẽ thở dài"->"c u4 j4+X E4+t_h 73+z a2 j2". Once the phones are analyzed, it will be split into a diphone or triphone depending on your needs, a diphone is a collection of two components, like the example above, we have diphone # -c or c -u4 ..., similar to triphone. The diphone will combine the physical parameters (if you want to change). Then, based on the trained data architecture, the system will synthesize and create the audio file as the text you enter.

One of the most important issues to address is the

standardization of Vietnamese text. The requirement is to solve common problems of standardizing documents in general and specific issues in Vietnamese. It is possible to deal with cases where the phonetic rules can not be applied to provide phonetic information. Specifically, acronyms (ĐH – Đại học; GS – Giáo sư; HN – Hà Nội), borrowings (taxi, scandal ...), numbers (numbers, phone numbers, address, time ...) , email addresses or URL addresses.

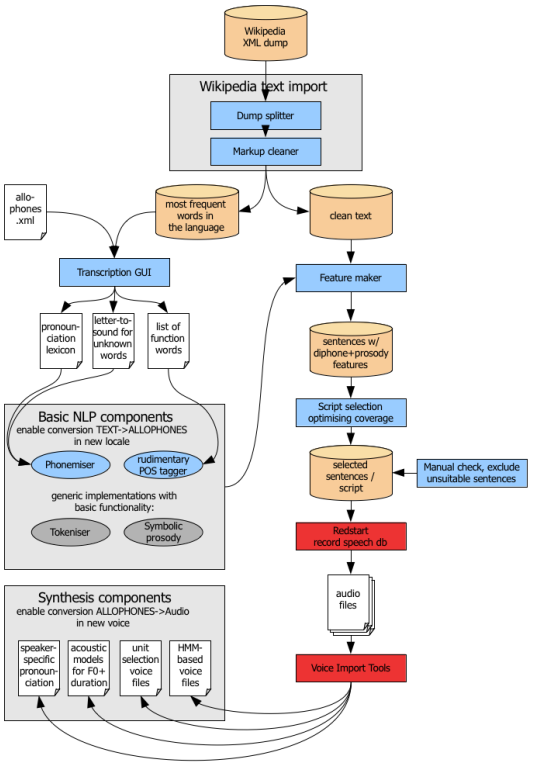


Figure 6 New language work flow

Processing has many complex problems because of its richness, diversity and local attributes. Many of the sub-categories used are pronounced in terms of basics that are quite different from the original pronunciation in the dictionary. In addition there is a high level of ambiguity in pronunciation, which makes more cases more pronounced, and the correct pronunciation depends on the context of the word. For example, "1994" can be read as " một chín chín bốn " or " một nghìn chín trăm chín mươi tư ". Acronyms are also ambiguous when an acronym has multiple applicants for the full word with it. For example: "HS" may be "Hồ sơ" or "Học sinh". This complexity requires consideration of the context element in the input text. And so the model needs to be able to read based on the context of these particular words.

In order to meet this requirement, a clear classification system for words that need to be standardized (NSW) should first be set up to deal with each type separately.

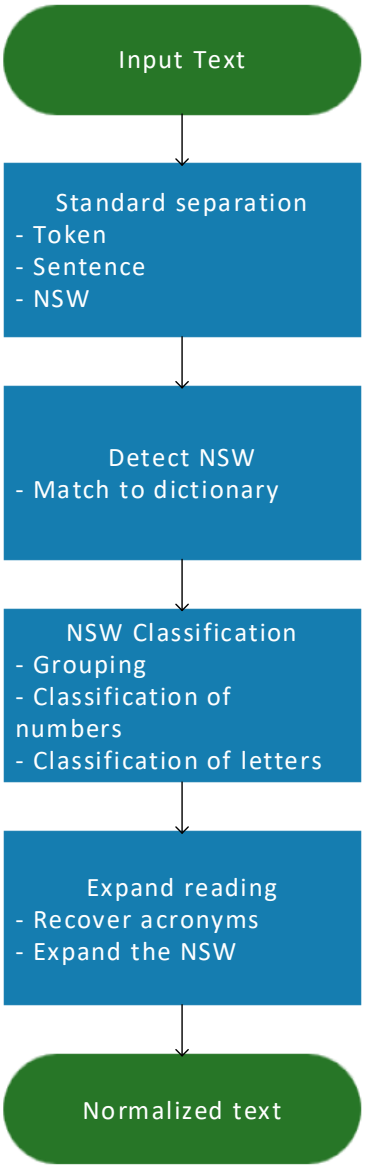


Figure 7 Vietnamese Standardized text block

The input text is first processed by the regular expression to identify the combinations of the numerical groups, after detecting the program group numbers that will invoke the corresponding processing functions to return the standard form. Then the extra white marks in the text are removed, with the white marks before and after the punctuation marks. Excess characters will be removed.

The next steps we follow the instructions of MaryTTS http://mary.dfki.de/index.html.

## 4. Evaluation

The objective of evaluation experiment is to show how the Vietnamese voice can be understand by Vietnamese native.

For testing, we take three comparative Vietnamese TTS systems:

- The vais[1] TTS system: which use HMM architecture, is a famous Vietnamese TTS system
- Google Text-to-Speech[2] is a screen reader application developed by Google for its Android operating system. It powers applications to read aloud (speak) the text on the screen which support many languages.
- The proposal system: as presented in section 2.

Testing data used for evaluation these systems which are 30 sentences in poems in the story of Nguyen Du, here are some examples:

- Trăm năm trong cõi người ta (One hundred years in the realms)
- Chữ tài chữ mệnh khéo là ghét nhau (Quite good words hate each other)

Voice output recorded 90 audio files, tested by 50 students studying at Geological University, 40 men and 10 women.

The test takers listen to each sentence three times, evaluating and scoring on a scale: 5 - Very good (like natural voice), 4 - Pretty (quite natural), 3-Moderate (Acceptable), 2 - Poor (Hard to hear), 1 - Bad (Inaudible).

The results are summarized as follows:

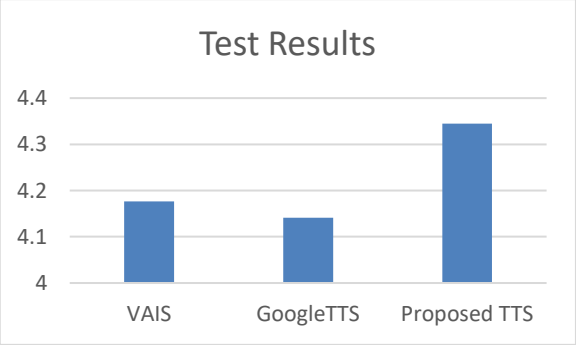


Figure 8 Test Results

The test results show that our pilot Vietnamese system has produced satisfactory results and opened up a great prospect for research investment to improve the quality of integrated voice is based on MaryTTS system.

## 5 Conclusion

This paper presents the experiment of building a TTS system for Vietnamese language, a primary language of Vietnam, using MaryTTS framework. The characteristics of Vietnamese language was studied for using to build the TTS system, thank to many linguistic researches. The Vietnamese synthesized speech was evaluated by Vietnamese speakers in perception tests. The result shows that the synthesized speech have a good quality, compare with other Vietnamese TTS systems. We can focus on further research to improve the quality of Vietnamese synthesis and also think about of quickly build Cambodia TTS through the use of MaryTTS.

## References

[1] J. Allen, M. S. Hunnicutt, D. H. Klatt, R. C. Armstrong, and D. B. Pisoni, *From text to speech: The MITalk system*. Cambridge University Press, 1987.

[2] F. Deprez, J. Odijk, and J. D. Moortel, "Introduction to multilingual corpus-based concatenative speech synthesis," in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[3] A. J. Hunt and A. W. Black, *Unit selection in a concatenative speech synthesis system using a large speech database*. 1996.

[4] T. S. Lê and H. Minh, "Nghiên cứu xây dựng phần mềm 'Tự động đọc văn bản chữ Việt' bằng phương pháp tổng hợp Formant," 2004.

[5] K. Tokuda, H. Zen, and A. W. Black, "An HMM-based speech synthesis system applied to English," in *IEEE Speech Synthesis Workshop*, 2002, pp. 227–230.

[6] T. T. Vu, M. C. Luong, and S. Nakamura, "An HMM-based Vietnamese speech synthesis system," in *Speech Database and Assessments, 2009 Oriental COCOSDA International Conference on*, 2009, pp. 116–121.

[7] J. Schmidhuber, "Deep learning in neural networks: An overview," *Neural Netw.*, vol. 61, pp. 85–117, 2015.

[8] C. Szegedy, A. Toshev, and D. Erhan, "Deep neural networks for object detection," in *Advances in neural information processing systems*, 2013, pp. 2553–2561.

[9] H. Ze, A. Senior, and M. Schuster, "Statistical parametric speech synthesis using deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*, 2013, pp. 7962–7966.

[10] T. T. T. Nguyen, "HMM-based Vietnamese Text-To-Speech: Prosodic Phrasing Modeling, Corpus Design System Design, and Evaluation," Paris 11, 2015.

[11] T. Do Dat, E. Castelli, J.-F. Serignat, T. Van Loan, and L. X. Hung, "Influence of F0 on Vietnamese syllable perception," in *9th European Conference on Speech*

[1] https://vais.vn/tong-hop-tieng-noi/

[2] https://translate.google.com/

*Communication and Technology (Interspeech 2005)*, 2005, pp. 1697–1700.
[12] D.-K. Mac, T.-L. Nguyen, A. Michaud, and D.-D. Tran, "Influences of speaker attitudes on glottalized tones: a study of two Vietnamese sentence-final particles," in *ICPhS XVIII (18th International Congress of Phonetic Sciences)*, 2015.
[13] "VTCC AI." [Online]. Available: https://vtcc.ai/tts. [Accessed: 04-Dec-2018].
[14] "Text to Speech – FPT Software – Powering Digital Transformation." .
[15] X. Huang, A. Acero, H.-W. Hon, and R. Foreword By-Reddy, *Spoken language processing: A guide to theory, algorithm, and system development*. Prentice hall PTR, 2001.
[16] P. Taylor, "Text-To-Speech Synthesis," *Camb. Univ. Press*, 2009.
[17] B. K. Lê and B. K. Lê, *Vietnamese-English, English-Vietnamese Dictionary: With a Supplement of New Words English-Vietnamese*. New York: Hippocrene Books, 1991.
[18] T. T. Đoàn, *Ngữ âm tiếng Việt*, Nhà xuất bản Đại học quốc gia Hà Nội. 1997.
[19] A. De Rhodes, "1651," *Dictionarium Annamiticum Lusit. Latinum*.
[20] H. Maspero, "Etudes sur la phonétique historique de la langue annamite. Les initiales," *Bull. LÉcole Fr. Extrême-orient*, vol. 12, no. 1, pp. 1–124, 1912.
[21] L. V. Lý, *Le parler vietnamien:(essai d'une grammaire vietnamienne):[sa structure phonologique et morphologique fonctionnelle]*. Huong Anh, 1948.
[22] M. B. Emeneau, *Studies in Vietnamese (Annamese) grammar*, vol. 8. University of California Press, 1951.
[23] L. C. Thompson, "A Vietnamese grammar.," 1965.
[24] L. C. Thompson, *A Vietnamese Reference Grammar*. University of Hawaii Press, 1987.
[25] J. Edmondson and N. V. Lợi, "Tones and voice quality in modern northern Vietnamese: instrumental case studies."," *Mon-Khmer Stud.*, vol. 28, 1997.
[26] D. X. Kien, "Re-consider a problem of Vietnamese phonetics: Syllable structure," *Hop Luu*, vol. 48, pp. 1–24, 1999.
[27] A. Michaud, "Final consonants and glottalization: new perspectives from Hanoi Vietnamese," *Phonetica*, vol. 61, no. 2–3, pp. 119–146, 2005.
[28] A. Michaud, T. Vu-Ngoc, A. Amelot, and B. Roubeau, "Nasal release, nasal finals and tonal contrasts in Hanoi Vietnamese: an aerodynamic experiment," *Mon-Khmer Stud.*, vol. 36, p. pp–121, 2006.
[29] A.-G. Haudricourt, "The origin of the peculiarities of the Vietnamese alphabet," *Mon-Khmer Stud.*, vol. 39, pp. 89–104, 2010.
[30] J. P. Kirby, "Vietnamese (Hanoi Vietnamese)," *J. Int. Phon. Assoc.*, vol. 41, no. 3, pp. 381–392, Dec. 2011.