VOS: the Corpus-Based Vietnamese Text-to-Speech System

Vo Quang Dieu Ha, Nguyen Manh Tuan, Cao Xuan Nam, Phạm Minh Nhut, Vu Hai Quan
University of Science, Vienam National University, Ho Chi Minh city
Email: vhquan@fit.hcmus.edu.vn

***Abstract*: This paper presents a complete specification of the Vietnamese speech synthesis system named VOS (Voice of Southern Vietnam). Due to the fact that current Vietnamese text-to-speech systems lack the naturalness of output synthetic speech, VOS is based on the unit selection approach which aims to achieve maximum naturalness. There are three main parts constituting VOS: a corpus manager, a synthesizer, and a transliteration model. Corpus manager manages automated speech indexing and segmentation for unit selection executed by the synthesizer, while transliteration model deals with the pronunciation of words in foreign languages. A comparative experimental evaluation of VnSpeech, VietVoice, and VOS is conducted using ITU-T P.85 standard. Results show that VOS outperforms the former two TTS systems.**

Key words:* *VOS, Vietnamese, speech synthesis, text-to-speech, corpus-based, unit selection

I. INTRODUCTION

Speech synthesis is a task of generating artificial utterances similar to human speech. This field of study is also known as text-to-speech (TTS) – the process of converting written text into speech. TTS systems have been studied and developed for different languages since 1968. There are four primary approaches in building a TTS system: concatenative synthesis, formant synthesis, articulatory synthesis, and statistical parametric synthesis. Various TTS systems have been recently developed for many languages including Japanese [5], Korean [6], Chinese [3], Thai [9], etc. In this paper, VOS (*Voice of Southern Vietnam*), a Vietnamese TTS system, is presented.

Vietnamese is a monosyllable, tonal language. Each word unit is pronounced as a syllable and its meaning depends on the tone. There are about 6596 phonetically distinguishable syllables [4] which comprise of legal combinations between basic syllables (i.e. syllables without tone) and five tones. Figure 1 illustrates the diacritics used for representing tones, including: level tone (denoted by "none"), high-rising tone (/), low-falling tone (\), dipping-rising tone (?), high-rising glottalized tone (~), and low glottalized tone (.). Although word, a group of one to several syllables, is the smallest syntactically meaningful unit, syllable is the basic pronunciation unit in Vietnamese speech. Thus, using syllable as a basic synthesis unit is an ideal choice for Vietnamese unit selection TTS systems.

Diacritic	none	/	\	**?**	~	**.**
Example	xa	xá	xà	xả	xã	xạ
Meaning	far	bow	snake	release	village	musk

Figure 1. Diacritics in Vietnamese.

Earlier developed Vietnamese TTS systems [7, 8, 11, 13] produce poor naturalness in output speech. The problems of these systems can be pointed out as: the robotic-sounding nature of formant-synthesis speech, and audible glitches in corpus-based synthetic speech which are caused by the simple technique of pre-recording /concatenating isolated syllables. In addition, the pronunciation of foreign words is very limited or not covered at all. Therefore, two main objectives for the development of VOS are:

1. to achieve maximum naturalness in output speech.

2. being able to pronounce any arbitrary foreign word.

To accomplish the first objective, VOS applies unit selection for the synthesis process with the use of advanced units (i.e. non-uniform units which can be either syllables, words, or phrases). Also for this objective, VOS aims to build a large speech corpus and segments it automatically using the speech recognition system that was reported in [10]. For the second objective, VOS employs a data-driven model in the text-normalization process.

In the remainder of this paper, Section 2 presents an overview of VOS system. Detailed expositions for VOS' primary components are undertaken in Sections 3-5 respectively. Section 6 focuses on experimental results and current developments of VOS. Finally, Section 7 gives conclusions and future works.

II. VOS' ARCHITECTURE

This section presents the general view of VOS' underlying structures. Figure 2 illustrates three main parts constituting VOS, and their interactions. The components are: the corpus manager, the synthesizer, and the transliteration model. Each one plays an indispensable role in the whole system. The corpus manager manages automated speech indexing and segmentation for unit selection executed by the synthesizer, while transliteration model deals with the pronunciation of words in foreign languages.

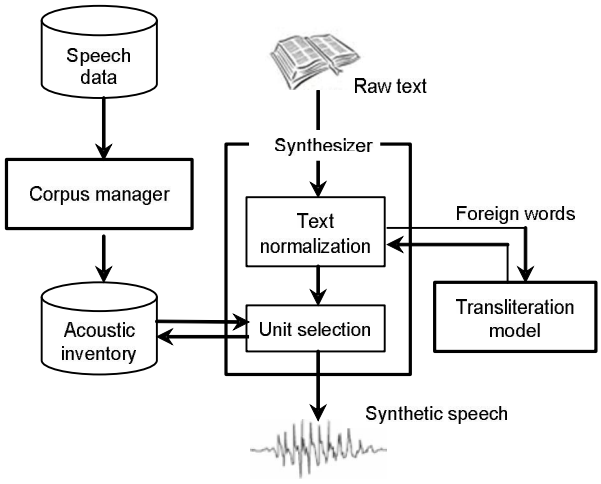


Figure 2. VOS system

Raw text input to the synthesizer will be first passing through the text normalization step. Numbers, symbols and abbreviations are expanded to their phonetic representations (i.e. syllables) using a set of rules and an abbreviation dictionary. Foreign words, if present, are sent to the transliteration model for retrieving their corresponding Vietnamese syllables. Normalized text from this step will be synthesized in the unit selection step. VOS employs unit selection to generate the most appropriate prosody for output speech. The whole unit selection process requires an acoustic inventory pre-built by the corpus manager.

III. CORPUS PREPARATION

The corpus manager segments pre-recorded utterances into syllables, words and common phrases by indexing their starting and ending offsets within speech files. This is done by running the forced-alignment mode of the speech recognizer which employs HMM-based acoustic models and a trigram language model. The result, consisting of only syllables, will be converted into words and common phrases using the unit inventory list.

A. Segmentation

The large vocabulary continuous speech recognition system in [10] is deployed for automated audio segmentation. The speech recognizer, constructed by HMM-based acoustic models and a trigram language model, is trained using the Vietnamese Broadcast News Corpus (VNBN) consisting of 27-hour speech gathered from numerous Saigon dialect broadcasters. Speech data were sampled at 16 kHz and 16 bits, and further parameterized into 12 dimensional MFCC, energies, and their delta and acceleration (39 length front-end parameters).

Having built the speech recognizer, forced-alignment mode is used for automated audio segmentation. The 41-hour speech corpus of one female speaker is built and then segmented to serve as the acoustic inventory for unit selection. The creation

of this corpus is conducted in two steps: text collection and speech recording. In the first step, text is collected with an effort to cover significant linguistic events (mainly on news), and phonetic criteria are taken into account aiming to collect text sentences that follow the desired distribution of syllables as well as prosodic features. Two criteria are used:

1. In addition to the requirement of phonetically rich sentences, all syllables must occur multiple times (from 15 to 50).

2. Different types of prosodic situations must be designed and incorporated into text as many as possible.

In the next step, collected text material will be brought into recording by a female speaker. Recorded utterances are then checked to verify speaking style, intensity level, script-consistency, etc. The resulting corpus consists of 41-hour speech sampled at 16 kHz and 16 bits (wav format) with a total size of 4.7GB.

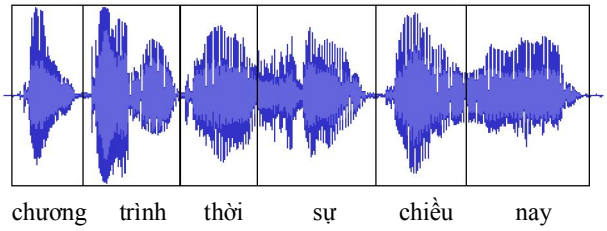


Figure 3. Syllable alignment

After building the speech corpus and applying forced-alignment process on it, all utterances will be segmented into syllables by indexing their starting and ending offset within speech files, as shown in Figure 3.

B. From syllables to words and phrases

In Vietnamese, a meaningful word consists of one to several syllables. Therefore, a unit selection TTS system that use only syllables as basic units cannot perform well in Vietnamese. Instead of that, VOS explores the use of advanced units which can be either syllables, words, or phrases. To make these advanced units available for use, syllable alignments must be converted into phrase alignments as shown in Figure 4. This is done by using the inventory list.

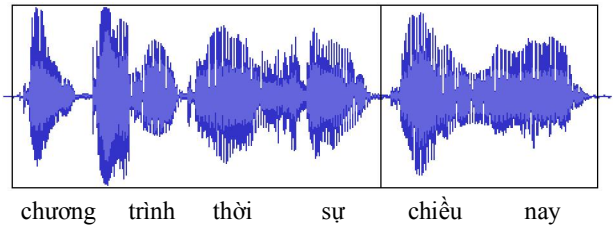


Figure 4. Phrase alignment

Basic units in the inventory list are constructed as follow. First, statistics on the 146M-word collection of newspaper text are done to select frequent units as shown in Table 1, in which words and phrases are treated in the form of syllable-sequences. These results are then filtered by the speech corpus' text (i.e. the prompts for recording), resulting in a list of 23092 words and common phrases. This list is then served as the unit inventory for the system.

Table 1. Statistics for selecting basic units

Common units	
# Syllables	**Count**
6	200994
5	173025
4	139167
3	90405
2	35455
1	2112

IV. SPEECH SYNTHESIS

A. Text normalization

Input text usually contains numbers, symbols, and abbreviations that must be expanded into phonetic representations before synthesizing. This process of expansion, called text normalization, is not a trivial task due to the ambiguous nature of numbers, symbols and abbreviations. For example, the number string "1-2" can be thought of as a football score or as a range from 1 to 2. VOS decides how to convert numbers/symbols based on a set of rules extracted from the 146M-word collection of newspaper text. Each rule describes a conversion method depending

on contexts of the sentence to which the numbers/symbols belong. A wide range of context types covered by VOS includes but not limited to football, addresses, date/time, grade, size, currencies, plain number, internet symbols, etc.

To deal with abbreviations, VOS employs the 3062-entry abbreviation dictionary. It is built from the same text corpus used to extract expansion rules for numbers/symbols. If an abbreviation stands for more than one complete form, its neighboring words (from the input text) will be taken into account to decide which form to use. In the worst case when an abbreviation is not found in the dictionary, VOS will try searching for its complete form right in the input text, and convert the abbreviation into its alphabetic form if failing.

Aside from numbers, symbols, and abbreviations, foreign words must also be converted to Vietnamese syllables. In VOS, this is done by the transliteration model which will be discussed further in Section 5.

B. Unit selection

Output from the normalization step is the normalized text containing only syllables, and hence there is no need for the text-to-syllable conversion step. The longest matching method [12] is applied from left to right to split each sentence into segments. A segment can be either a phrase (highest priority), a word, or a single syllable (lowest priority) available in the inventory. These segments are then put into unit selection. VOS employs unit selection to generate the most appropriate prosody for output speech. Each unit (i.e. segment) is selected from the inventory with regard to its left and right units to match the prosody.

The problem can be formally stated as follows. Let $S = \langle s_1, s_2, ..., s_I \rangle$ be the sentence being processed, with s_i being a syllable within S, and $U = \langle u_1, u_2, ..., u_M \rangle$ be the set of all candidate units available for selection. $R(u_t, P_t)$ is defined as the selection cost of unit u_t for segment $P_t = (s_i, s_{i+1}, ..., s_{i+L-1})$ (i.e. a phrase of length L), and $J(u_t, u_{t+1})$ is defined as the concatenation cost of u_t and u_{t+1}. The total cost for synthesizing S is then given by:

$$C(U,S) = \sum_{t=1}^{T} R(u_t, P_t) + \sum_{t=1}^{T-1} J(u_t, u_{t+1}) \qquad (1)$$

where T is the number of segments into which S is split. Candidate units must be chosen in such a way that (1) is minimized. This optimization problem can be viewed as a graph search problem and can be solved using A-star (A*) algorithm. The whole set of all candidate units is organized to construct a weighted directed graph with each vertex corresponding to a candidate unit. Vertices are connected with direction in the time-line order (the order in which they appear in the sentence). Weight of the edge connecting u_t to u_{t+1} is given by the sum of $J(u_t, u_{t+1})$ and $R(u_{t+1}, P_{t+1})$. Figure 5 illustrates the detailed A-star algorithm for unit selection in VOS.

1. Initialization:
   - set S = <root>, b(S) = 0, f(S) = 0, pn(S) = null
     (pn(S): a parent node leading to S)
   - put {S, f(S), b(S), pn(S)} in the **Open** priority queue
     (the lower f(n) for a given node n, the higher its priority)
   - create an empty list: **Close** = {null}
2. Begin loop
3. Dequeue the first node N from **Open** and put it into **Close**.
4. If N is a candidate unit for the last segment, exit algorithm with the solution obtained by back-tracking from N to S using pn(N).
5. Expand the paths to all vertices leading from N. Let E(N) be the set of these vertices.
6. For each V in E(N):
6.1. set h(V) = the number of remaining syllables after V
     b(V) = b(N) + J(N,V) + R(V)
     f(V) = b(V) + h(V)
     pn(V) = N
6.2. ***if*** V ∈ **Open** and b(V) < Open.b(V) then update:
   Open.f(V) = f(V), Open.b(V) = b(V), Open.pn(V)= N
   else if V ∈ **Close** and b(V) < Close.b(V) then:
     set Close.pn(V) = N, and adjust g, f for all the paths containing V.
   else
     put {V, f(V), b(V), pn(V)} into **Open**.
7. Go to step 2.

Figure 5. A algorithm for unit selection in VOS*

The path decision function for node u_t is given by:

$$f(u_t) = b(u_t) + h(u_t) \quad (2)$$

where $b(u_t)$ is the total cost for synthesizing up to u_t (i.e. the cost for synthesizing $(P_1, P_2, \ldots, P_t)$) and can be calculated inductively as:

$$b(u_t) = b(u_{t-1}) + J(u_{t-1}, u_t) + R(u_t, P_t) \quad (3)$$

The heuristic function $h(u_t)$, for a given node u_t, is indeed the number of remaining syllables after selecting that node. It is worth noting that the concatenation cost is intended to minimize audible signal discontinuities between two successive acoustic units. In VOS, it is given by the weighted sum of three components: the Euclidean distance between MFCC d_M, the absolute differences in log power d_L and pitch d_P:

$$J(u_t, u_{t+1}) = w_M d_M(u_t, u_{t+1}) + w_L d_L(u_t, u_{t+1}) + w_P d_P(u_t, u_{t+1}) \quad (4)$$

where w_M, w_L, and w_P are the corresponding weights of each component. These weights are chosen during the corpus preparation phase and are optimized using a trial-and-error procedure. VOS aims to minimize discontinuities at concatenation points by minimizing the concatenation costs without applying smoothing techniques. Furthermore, intonations are not modified.

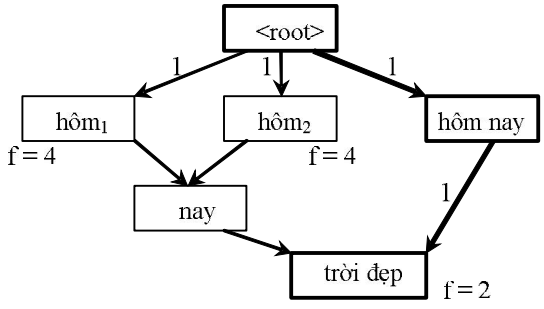


Figure 6. An example of A algorithm for unit selection*

A concrete example for selecting best candidate units for the sentence "hôm nay trời đẹp" using A* algorithm is depicted in Figure 6. The optimizing path (<root> → "hôm nay" → "trời đẹp") is chosen based on its overwhelming decision functions (f = 3 and f = 2 for each node).

V. CROSS-LINGUAL PHONEMIC TRANSLITERATION

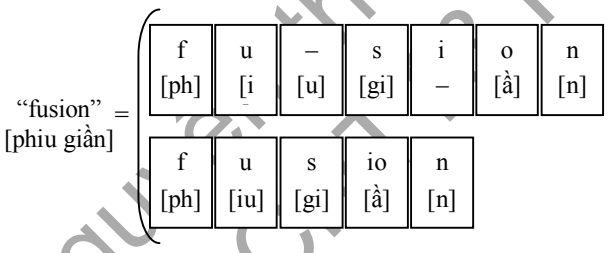
A significant problem of Vietnamese text-to-speech systems is the pronunciation of words in foreign languages. A common solution to this problem is to make use of a transcription dictionary. Despite its effectiveness, this approach has serious limitations: making a cross-lingual pronunciation dictionary of large size by hand is costly and requires a lot of effort. Furthermore, the number of available entries is finite and therefore not flexible because TTS systems are expected to handle arbitrary words. Alternatively, data-driven approaches can be employed to overcome these limitations by learning samples and predicting unseen words. In VOS, joint-sequence model [2], a data-driven approach, is applied to transliterate foreign words into Vietnamese syllables.

A. Cross-lingual joint-sequence model

The fundamental idea of joint-sequence model is based on the concept of graphone [2], a joint unit between graphemes (letters) and phonemes. Let G be the set of foreign graphemes and Φ be the set of Vietnamese phonemes, then a cross-lingual graphone q is a pair of a grapheme sequence $g \in G^*$ and a phoneme sequence $\varphi \in \Phi^*$ of possibly different length:

$$q = (g, \varphi) \in Q \subseteq G^* \times \Phi^* \quad (5)$$

Here G* and Φ* denote the set of all strings over symbols in G and Φ respectively (Kleene star). In the assumption of joint-sequence model, each word and its pronunciation are generated by a common sequence of graphones, but the number of possible graphone sequences varies depending on the ways of segmentation. For instance, the word "fusion" and its pronunciation can be represented by one of the following graphone sequences:



The generation of graphone sequences is under-run by standard N-gram probabilities:

$$p(q_1^K) \approx \prod_{i=1}^{K} p(q_i \mid q_{i-1},...,q_{i-N+1}) \qquad (6)$$

where $q = q_1, q_2, \ldots, q_K$ is a graphone sequence of length K. The graphone inventory and its accompanied N-gram parameters can be estimated from the training data using discounted EM algorithm [2].

B. Transliteration

Given a word in an orthographic form (i.e. a sequence of graphemes) its corresponding pronunciation is the most likely phoneme sequence yield by Bayes decision rule:

$$\varphi = \arg\max_{\varphi' \in \Phi^*} p(g, \varphi') \qquad (7)$$

Having estimated model parameters, (7) can be approximated in the form of:

$$\varphi(g) = \varphi(\arg\max_{q \in Q^* | g(q)=g} p(q)) \qquad (8)$$

That is, searching for the most likely graphone sequence which matches the same spelling as given, and then project it into phonemes. The resulting phonemes are then assembled into Vietnamese syllables for unit selection.

In VOS, the model is trained with a cross-lingual pronunciation dictionary of 5K foreign words, and tested with another 2K foreign words. Word accuracy rate (WAR) reaches 82.89%.

VI. EXPERIMENTS & DEVELOPMENTS

A. Evaluation metrics

Table 2. ITU-T P.85 rating scales used in evaluating VOS

Rating scale	Interpretation
Overall impression	"How do you rate the overall quality of the sound?"
Comprehension	"Did you find certain words hard to understand?"
Articulation	"Were the sounds distinguishable?"
Pronunciation	"Did you notice any anomalies in the pronunciation?"
Voice pleasantness	"How would you describe the voice?"

A challenge in evaluating speech synthesis systems is the lack of universally agreed objective evaluation criteria and data. However, in [1] the ITU-T P.85 standard was verified to be reliable and thus is used as an evaluation scheme for VOS. Listeners are presented with synthetic speech and are asked to rate on five scales: overall impression, comprehension, articulation, pronunciation, and voice pleasantness. Each of these scales ranges from 1 (poor) to 5 (good). Table 2 shows the rating scales and their corresponding interpretations.

B. Comparative evaluation

Table 3. Rating scores

Rating scale	VnSpeech	VietVoice	VOS
Overall impression	2.68 ± 0.13	3.03 ± 0.17	4.34 ± 0.23
Comprehension	2.43 ± 0.21	3.28 ± 0.12	4.82 ± 0.06
Articulation	3.12 ± 0.18	2.9 ± 0.19	4.42 ± 0.17
Pronunciation	2.12 ± 0.22	3.14 ± 0.08	4.16 ± 0.11
Voice pleasantness	2.43 ± 0.15	2.98 ± 0.23	3.95 ± 0.05

In this experiment, a comparative evaluation is carried out between VOS and two earlier Vietnamese TTS systems: VnSpeech [7] and VietVoice [8]. A set of 200 sentences is put in line for synthesis using each system respectively and then presented to 10 native listeners. The listeners are asked to rate output speech

of each system on ITU-T P.85 rating scales. Average rating scores and their accompanied standard deviations are shown in Table 3. Furthermore, Figure 7 plots 95% confidence intervals for the ratings of each system. Obviously, VOS outperforms VnSpeech and VietVoice in all aspects.

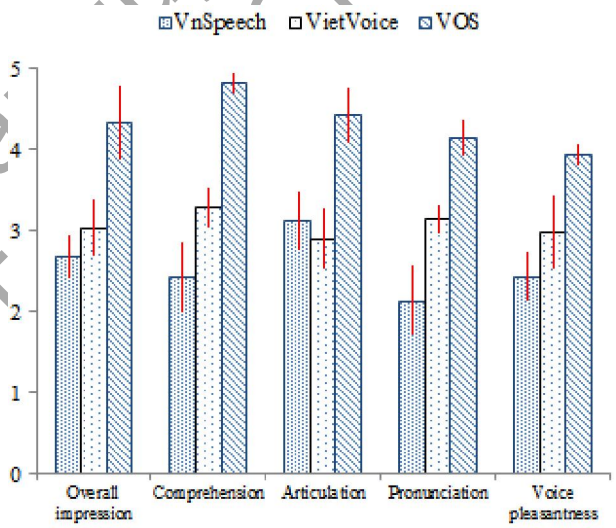


Figure 7. Systems' performances.

C. Developments

Three versions of VOS are developed on different platforms: a standalone synthesizer for Windows, a compact plug-in for web browsers, and a mobile version for iPhone. All these versions are developed based on the VOS core library which is a concrete implementation of the methods described in this paper. Therefore, it is easily extended and deployed to any new platform or application. VOS is also available for testing at http://www.ailab.hcmus.edu.vn/slp/vos.

VII. CONCLUSION

A complete specification of VOS is presented in this paper. Utilizing unit selection approach, VOS provides natural and intelligible synthetic speech. Aside from the great naturalness, VOS also offers the capability of synthesizing any arbitrary foreign word which is very limited or not covered at all in the other Vietnamese TTS systems. Having achieved its two main objectives, VOS is put into rapid developments and aims for supporting various real-life applications. However, as unit selection synthesis requires a large storage of speech databases, coding algorithms will be considered for compacting the data in future works. Furthermore, intonation and rhythm of the synthetic speech will also be taken into account.

REFERENCES

[1] Y. Alvarez, M. Huckvale, "The reliability of the ITU-T P.85 standard for the evaluation of text-to-speech systems," Proceedings of the ICSLP'02, Denver, pp. 329-332, 2002.

[2] M. Bisani, H. Ney, "Joint-sequence models for grapheme-to-phoneme conversion", Speech Communication, vol. 50, no. 5, pp. 434-451, May 2008.

[3] M. Dong, K.T. Lua, H. Li, "A unit selection-based speech synthesis approach for Mandarin Chinese," Journal of Chinese Language and Computing, vol. 16, no. 3, pp. 135-144, September 2006.

[4] P. Hoang, Syllable Dictionary, Danang Publishing House, 1996.

[5] H. Kawai, T. Toda, J. Yamagishi, T. Hirai, J. Ni, N. Nishizawa, M. Tsuzaki, K. Tokuda, "XIMERA: a concatenative speech synthesis system with large scale corpora," IEICE Trans. J89-D-II, no. 12, pp. 2688-2698, December 2006.

[6] S.J. Kim, J.J. Kim, M.S. Hahn, "Implementation and evaluation of an HMM-based Korean speech synthesis system," IEICE Transactions on Information and Systems, vol. E89-D, no. 3, pp. 1116-1119, 2006.

[7] H.M. Le, VnSpeech, http://www.vnisg.com, 2009.

[8] T.H. Le, VietVoice, Virtual Voice Inc., http://noitiengviet.ca, 2009.

[9] L. Narupiyakul, A. Khumya, B. Sirinaovakul, N. Cercone, "A stochastic knowledge-based Thai text-to-speech system," Mathematical and Computer Modeling, vol. 42, issues 1-2, pp. 1-16, July 2005.

[10] T. Nguyen, Q. Vu, "Advances in acoustic modeling for Vietnamese LVCSR," Proceedings of the IALP'09, Singapore, pp. 280-285, 2009

[11] T.N. Pham, Vietnamese Voice, http://sourceforge.net/projects/vietnamesevoice, 2009.

[12] T. Theeramunkong, S. Usanavasin, "Non-dictionary-based Thai word segmentation using decision trees," Proceedings of the First International Conference on Human Language Technology Research, San Diego, pp. 1-5, 2001.

[13] T.T. Vu, M.C. Luong, S. Nakamura, "An HMM-based Vietnamese speech synthesis system," Proceeding of 12th Oriental COCOSDA, Tsinghua University, pp. 108-113, 2009.

AUTHORS' BIOGRAPHIES

Ha D. Q. Vo received the B.S. degrees in Information Technology, in 2009 from the University of Science, VNU-HCM. Currently, she is a research assistant in the Artificial Intelligent Laboratory, at which she's doing research on speech recognition and speech synthesis. She also works as a teaching assistant at the Department of Knowledge Engineering, University of Science, VNU-HCM.

Tuan M. Nguyen received the B.S. degrees in Information Technology, honor programs, in 2008 from the University of Science, VNU-HCM. Currently, he is a research assistant in the Artificial Intelligent Laboratory, at which he's doing research on speech recognition and speech synthesis. He also works as a teaching assistant at the Department of Knowledge Engineering, University of Science, VNU-HCM.

Nam X. Cao received the B.S. and M.S. degrees in Computer Science from University of Science in 2007 and 2010. From 2003 to 2007, he was a student in Faculty of Information Technology of University of Sciences. Currently, he is a teaching assistant at the Knowledge Engineering Department. Aside from that, he is also a member of the Artificial Intelligent Laboratory, where his works focus on the problems relating to speech synthesis and speech recognition.

Nhut M. Pham received the B.S. and M.S. degrees in computer science from the University of Science in 2006 and 2009, respectively. Since 2009, he has been working as a research assistant at the Artificial intelligent laboratory (AI-Lab), where his works focus on the fields of spoken language processing.

Quan H. Vu received the Ph.D. degree in computer science from the University of Trento, Italy in 2005. He is currenlty the Head of the Artificial intelligent laboratory (AI-Lab), where his works focus on the fields of spoken language processing.