International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-8 Issue-11, September 2019

A Method for Automatic Vietnamese Speech Segmentation

Tuan Tran Anh, Mong Nguyen Huu, Khanh Nguyen Trong

Abstract: ***In speech synthesis and recognition, the segmentation is an important step. The result of further steps depend completely on this process. There are several effective segmentation method in the literature, but for Vietnamese speech, researchers usually base on their experience to set the length while using sliding window. It causes an inefficient segmentation; and they need to try with the other value (length of voice). In this paper, we propose a method supporting in segmentation for Vietnamese speech and automatically determine the suitable length of voices and silent pause. We firstly estimate, by experimenting, the min and average length of a voice and a silent pause for Vietnamese speech in three main type speaking (slow, normal and fast). Then, based on these values, we start to segment the voice and pause by sliding window with proposed algorithm. Experiment results show that the proposed method can be used to effectively segment the Vietnamese speech.***

Keywords: Segmentation, automatic speech segmentation, Vietnamese speech segmentation.

I. INTRODUCTION

Automatic speech synthesis and recognition is a challenging task that has attracted a lot of attention from research community for two decades. It aims to receive linguistic messages from speech by describing signals in the form of separate units. The separation process is usually based on segmentation tasks where speech signals are segmented based on their features. One of the most important problem of speech segmentation is how to find word boundaries in spoken language when the underlying vocabulary is still unknown.

According to [1], different speakers and different languages can affect the segmentation process. Because, the speech is characterized by the different types of intonation or rhythm created by changes in pitch and sub-glottal pressure, as well as by different sounds of the language. Therefore, these characteristics should be considered in extracting speaker and language information. For several languages, the issue becomes an important challenge, like Vietnamese.

Manuscript published on 30 September 2019.
***Correspondence Author(s)
Phd. Student. Tuan Anh Tran, Department of Computer Science, Faculty of Information Technology, Industrial College, in Thanh Hoa, Vietnam, (84 4) 819801201, (e-mail: tuankhhtqt@gmail.com).
Dr. Mong Nguyen Huu, Department of Computer Science, Department of Information Technology, Military Technical Institute, in HaNoi, Vietnam, (84 4) 913002799, (email: nghm06@yahoo.com).
Dr. Khanh Nguyen Trong*, Software Engineering, Posts and Telecommunications Institute of Technology, Hanoi, Vietnam, (84 4) 912314482., (e-mail: khanhnt@ptit.edu.vn); Sorbonne University, IRD, UMMISCO, JEAI WARM, F-93143, Bondy, France

The Vietnamese language is predominantly a monosyllabic language in which majority of words have one syllable [2]. The language also have some disyllabic and polysyllabic words. A syllable is usually repeated with or without variation of either one sound or a tonal aspect.

Vietnamese language has 15 punctuation marks with different meanings. They are used to segment sentences; but they are not pronounced while speaking. Therefore, they are only implicitly existed in the positions of silence. An essential acoustic feature of every language is the intonation and rhythm. These features can be varied from one to another, or according to the context and also to the punctuation marks. Therefore, when speaking or reading, even with a same punctuation mark, its correspond silence is different. Thus, to determine a silence corresponding to which punctuation marks is a difficult task. For Vietnamese speech, to identify the punctuation marks, there are two approaches: (i) based on acoustic features of Vietnamese speech and (ii) based on Vietnamese language grammar.

Recently, there have been many researches interested in the field of speech segmentation. In general, the authors based on two main feature of audio: (i) time-domain features and (ii) frequency-domain features. The first one are widely used for speech segment extractions. These features are useful when it is needed to have algorithm with simple implementation and efficient calculation. For the second one, to extract frequency-domain features, discrete Fourier transform can be used. The Fourier representation of a signal demonstrates the spectral composition of the signal. The most information of speech is amassed in 250Hz-6800Hz frequency range.

The segmentation is an important step in the whole process of speech synthesis and recognition. The result and the accuracy of such process depends completely on it. Actually, for Vietnamese voice, to segment a voice, many researchers usually base on theirs experience to set the length of voice (or silent pause). It causes, in many cases, an inefficient segmentation; and they need to try with the other value (length of voice).

Therefore, our objective thus is to develop a method for Vietnamese speech segmentation. We will based on frequency-domain features of recorded speech, and then identify the correspond punctuation mark. Thus, the speech will be well segmented.

This paper is organized as follow: Section 2 presents the related works in the; Section 3 presents the way we extract data from raw speech; our experiments will be introduced in the Section 4; Section 5 is our conclusions and future works.

II. RELATED WORK

In general, speech segmentation can be grouped in various perspectives, but in general it can be divided into two group: supervised and unsupervised methods [3]. The first one relates to methods that based on existing labeled data or a priori knowledge, while the second one require no training data for segmenting the speech signal. Instead, they use sets of rules derived from or encoding human knowledge to segment speech.

In regards to the supervised methods, usually, an orthographic or phonetic transcription is used as a parallel contribution with the speech, or training the algorithm with such data. The approaches are generally computationally consuming. This group includes algorithms using recognition with Hidden Markov Models (HMMs) [4], Dynamic Time Warping (DTW) [5] or Artificial Neural Networks (ANNs) [6]. The algorithms of this group are employed only in the training stage.

For instance Zioko *et al.* [5] used the wavelet method to recognize the beginnings and ends of the phonemes based on (DWT) analysis and with the aid of spectral analysis, it happened to be an extremely effective technique.

Sharma *et al.* [7] presented a method for automatic segmentation of continuous speech signal of Punjabi language. The authors found that group delay function is better representation of STE for syllable boundary detection.

Nagarajan *et al.* [8] used a minimum phase group delay based approach to segment spontaneous speech into syllable like units. They analyzed over 5000 speech dialogs in Tamil language. The duration of the speech signal varies from 0.5 sec to 25 sec and obtained good result.

Prasad *et al.* [9] proposed a novel approach for segmenting the speech signal into syllable-like units. They propose a group delay based approach for processing the short-term energy to determine segment boundaries. The performance of this technique was tested on both continuous speech utterances and connected digit sequences. The error of segmenting boundary is less than 20% of syllable duration for 70% of the syllables.

Malcangi [10] demonstrated soft computing method for segmentation of speeches into phonetic units. The author used fuzzy decision logic to infer phonetic unit separation point and proved that this approach is more effective in segmentation of phonetic units.

Rahman *et al.* [11] used feature extraction approaches for segmenting Bangla speech sentences into word or sub-words. They proposed a method based on two speech feature namely time domain and frequency domain features and achieved segmentation accuracy of 96%.

Jayasankar *et. al* [12] proposed an algorithm for automatic segmentation of Tamil voiced speech. They use absolute energy and zero crossing rate to process speech samples to accomplish the segmentation. Chaudhury *et. al* [13] presented a word separation algorithm for Real Time Speech. The algorithm was developed by considering prosodic features with energy. The proposed algorithm correctly detects word boundaries of 98%. Stolcke *et al.* [14] have presented techniques for boosting the accuracy of automatic phonetic segmentation based on HMM acoustic-phonetic models. They have shown improved test results by using more powerful statistical models for boundary correction that are conditioned on phonetic context and duration features. They concluded that combining multiple acoustic front-ends gave an additional gain in accuracy, and that conditioning the combiner on phonetic context and side information helps with results, which reduced segmentation errors on the TIMIT corpus by almost one half, from 93.9% to 96.8% boundary accuracy with a 20-ms tolerance.

For the unsupervised methods, it usually relates to problem where the background knowledge about the phonetic content is lacked and are then based predominantly on statistical signal analysis. Due to this, the first phase of those methods relies completely on the acoustical features present in the signal. The second phase or bottom-up processing is normally built on a front-end parameterization of the speech signal, often using MFCC, LP-coefficients, or pure FFT spectrum.

Machine learning and neural network based methods like CNN and DNN are the approaches that inspire recently many researches in the field, for both supervised and unsupervised methods [15,16]. For example, in the work of Kamarauskas *et al.* [17], the authors based on a technique using perceptron and back propagation artificial neural network to recognize distinctive phonemes, by utilizing various features of the speech signal used in speech or speaker recognition (linear prediction coding, Cepstral coefficients, and coefficients of the Fourier transform). Similarly, Amirgaliyev *et al.* [18] presented a segmentation approach that based on co-functional modality, namely, a signal from a human's vocal cords vibrations. In a general sense, here modality refers to the mode of existence of any object, or the course of any phenomenon, or a mode of understanding, assertion about the object, phenomenon, or event.

In Vietnam, there are also some researches on Vietnamese speech segmentation. For instance, Nguyen *et al.* [19] presented a Vietnamese speech synthesis system based on deep neural networks. The system achieves superior MOS (Mean Opinion Score) of intelligibility and naturalness, compared to other Vietnamese TTS systems. For example, Ngo *et al.* [20] applied two approaches: (i) machine machine learning approach using maximum entropy (ME) and conditional random fields (CRFs); (ii) deep learning approach using bidirectional Long Short-Term Memory (LSTM) with a CRF layer (Bi-LSTM-CRF) for automatic dialog act segmentation. This is the first work that applies deep learning based approach to dialog act segmentation.

III. METHODOLOGY

A. *Pre-processing*

Regarding to the acoustic aspect, voice is a sequence of continuous sounds (syllables) where the silent pause is situated at between two adjacent sounds. The pause time is usually shorter than the voice. However, at the end of a sentence or at the end of a speech, when speaker pause, the time is relatively long, sometimes much longer than the voice.

On the audio side, the boundary between the two sounds is not clear: the sound of a sentence is often continuous and consecutive. The problem is that when speaking, the repercussion of each word sound is usually stretched, so that the listener feels like a sequence of continuous sound. This phenomenon is illustrated by the graph in Figure 1.

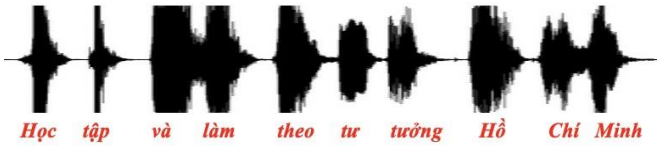


Fig. 1. Vietnamese speech example and its frequency-domain features

According to the graph as a kind of data, audio can be represented by the amplitude (sound intensity), thus it is difficult to verify if the sound is the voice and if the sound is silence. However, we always know that a sentence is a series of sounds consisting of continuous sounds with the silence between them. If speaking in an absolutely quiet environment and each voice is well separated, then pauses will be clearly identified: the pauses are the blank segments on the horizontal axis where data is segmented with the size of samples equal to zero (Figure 2).

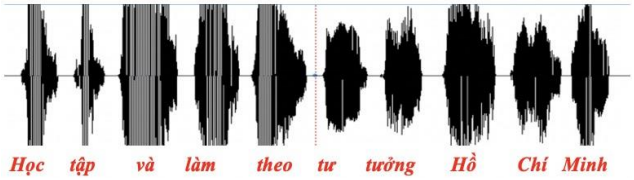


Fig. 2. The speech example after processing

In the ideal environment, as shown in Figure 2, the voice and pause are differentiated by their amplitude: pauses are the continuous graph with a level of zero. Therefore, if the pause and the voice can be separated clearly, we can eliminate the noise at the pause; and thus we can the speech that contains only the voice and the absolute pause. Such speech will support an excellent input for the other task, like segmentation or recognition.

In fact, if some small pieces at the starting and ending point of voice samples are removed, acoustically the change is not really clearly. Therefore, with a suitable threshold, we can separate the voices with a concrete distance that based on their position and component in the sentence.

In order to calculate these distance, we have done an experiment where 50 participants was requested to read a same text many times, with different speeds. All speeches are recorded, and also we can calculate the time for each voice and silence in each speech. Firstly, we observed the following results:

- Whether the speech is fast or slow, the average length of each voices is quite the same.
- The speed of speaking depends only on the length of the pause: the slower we speak, the higher pause is, vice versa the faster we speak, the smaller silence is.
- The above results do not depend on the sampling frequency when recording.

To compute the time of each voices and pause, we perform as following. Let Δt be the sampling time and n be the number of voice samples (or also the sample silence), Then, the length of the voice (or pause) is $n\,\Delta t$. Δt is calculated according to the formula:

$$\Delta t = \frac{1}{f_{\text{mau}}}, \tag{1}$$

wherein, f_{mau} is the sampling frequency. The f_{mau} frequency is included in the header of the wave file. The voice or pause length is calculated by the formula:

$$t = \frac{n}{f_{\text{mau}}}. \tag{2}$$

Accordingly, we obtain the following results:

Table- I. The average voice and silence length in second

Minimal length	Slow speed	Normal speed	Fast speed
Voice	0,257959	0,256304	0,251655
Silence	0,071882	0,032154	0,025555

We also need to emphasize that the length of the voice and the pause are not dependent on the sampling frequency. To illustrate that we did another experiment as follows: recording a speech three times with the frequency of 8000Hz, 44100Hz, 192000 Hz and we get the following results:

Table- II. The average voice and silence with different recording frequency

	8000Hz	44100Hz	192000Hz
Average voice length	0,255364	0,256114	0,256759
Average silence length	0,081397	0,085674	0,087115

The smallest lengths of voices and pauses take an important role in speech segmentation. By this value, we can determine the smallest threshold. Thus, we have a limitation for further segmentation operations. In the next sections, this value will be used to segment the voices.

B. *Framing*

In general, the sound signal is not stationary; moreover, the data size of speech is quite long, much longer that the correspond text. We cannot apply directly the segmentation to the whole speech at time. Therefore, instead of analysis the whole sound data, we segment the speech into short snippets of audio or frames. Because, at a given, very short time frame (10-50ms), the speech segment is close to stationary. In this interval, speech signal remain unchanged.

We consider the speech is a sequence of frames. Each frame is a piece of data, including frame numbers and its amplitude. The sound signals are converted then from analog to digital by getting divided into small frames between 20 and 40 ms during the framing stage and it is divided into N frames. The sound signal is moved by sliding the sound signal at the windowing stage.

As a result, the feature of sound is now present by those frames. We consider an important feature of the frame that is "absolute weight" which represent the quantity of the frame. Suppose the frame is denoted as ***KH***, then |*KH*| denoted as the "absolute weight" of the frame and is calculated according to the following formula:

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

$$|KH| = \sum_{i=1}^{k} |KH(i)|, \quad (3)$$

where, frame KH is a one-dimensional array with k elements. We see that, $|KH| \triangle t$ is the area of the region created by the frame in the graphs.

Based on this feature we can distinguish KH is the frame of the voice or of the pause. Because, in an ideal recording environment without noise, the pauses are spaces that mean the sample values are zero and $|KH|$ of frames also equal zero. If KH belongs to the voice, then it is definitely different from 0. In the general case, when the recording environment is noisy, $|KH|$ of the two frames are also different: $|KH|$ of the voice frame is always greater than that of pauses. The greatest $|KH|$ value of the pause frame also depends on the different types of speeches: loud speaking, speaking, normal speaking quiet speaking and so on

For our research, we are interesting in normal speeches where their power ranges from 25dB to 35dB, with the number of frame not greater than 100. By experimented with the 50 participants, we found that: $min|KH|$ of the voices is always higher than $max|KH|$ of the pause. The border frames are not much different, but $|KH|$ of the voice is still higher than $|KH|$ of the pause. If the recording environment is absolutely clear and the pronunciation is standard, $|KH|$ of the pause is always equal to 0, and in other cases this value is different from 0, but always smaller than $|KH|$ of the voice. These values are not fixed and vary according to speech intensity and noise.

Let set the length of a frame to the number of its samples (greater than 2), we will find a minimum value for this length in order to distinguish voice and silent pause. According to the previous experiment, for a normal voice recorded with frequency of 44100 Hz, the acceptable length of KH length is 50.

Let kh_{min} be the minimum length of $|KH|$, if KH has $|KH|$ less than or equal kh_{min}, it will be a silent pause; and it will not if vice versa. If these voices in KH, $|KH|$ will greater than kh_{min}. Therefore, the accuracy of voice segmentation depends on the sample number of KH.

We have already performed an experiment with 50 participants to find $|KH|$ of voice and also of silent pause: With a consecutive voice and silent pause, we take 3 KH and calculate the average $|KH|$. The result of the experiment is illustrated in Table 3. Based on these values, we propose a method to segment Vietnamese voice, as presented in the next section.

Table- III. The average |KH|

KH	$\|KH\|$ of voice	$\|KH\|$ of pause
Beginning	0,59708	0,04019
Middle	2,09732	0,05548
End	0,58420	0,03708

C. *Algorithm*

The objective of our algorithm is to segment voices and pause, and also calculate their time length. The minimum length is used for segmentation, while the average length is used to evaluate the result.

In order to do this, from X, we need to separate voices, pauses and compute their length. Moreover, we also update the minimum and the average length.

Suppose we have a speech which was recorded to a wave file. The data then is pre-processed by cutting the gaps before and after each voice. We obtain a set X that is a series of amplitude of the speech.

Let *ddt* be the array of voice length, *ddl* be the one of pause, *M* be the number of these ddt array (that mean *M-1* is the number of *ddl* array). Let *nt* be the sample number of voices or silence; *k* be the kth of the array *T*, *ddt*, *ddl* with $1 \le k \le M$.

Consider a speech with a certain rhythm, as presented in Table 1. Let ddt_{min} and ddl_{min} be the minimum value of voice and pause length; and let NTmin and NLmin be the minimum number of samples of voices and pause, we have:

$$NT\min = ddt\min * SR; \quad NL\min = ddl\min * SR, \quad (4)$$

in which, SR is the sampling frequency.

We consider the frame that has *ktm* samples and $|KH| = khl$. The value of *khl* is computed as presented in the previous section. Based from this suppose, we propose the following algorithm in order to segment the voice:

```
1. BEGIN
//Take the input by sliding the window with
//the minmum as shown in Table 1, for 3 type
//of speech
 2. seg_voice = []
 3. seg_pause = []
 4. X = []
 5. voice_length = 0
 6. Sliding_window_size = min_voice_len
 7. input_speech = read_wave_file();
 8. X = Sliding_window(input_speech);
 9. for i=0 to length(X):
   10. IF length(X) - i > |KH| then:
     11. KH = X[i];
   12. ELSE then:
     13. seg_voice += X[i]
     14. voice_len=compute_len_seg(seg_vo
         ice)
     15. continue
   16. IF KH == voice then:
     17. seg_voice [i] += KH
     18. continue
   19. ELSE:
     20. voice_len=compute_len_seg(seg_vo
         ice[i)
    //check EnOffFile
     21. seg_pause[i] += KH
    //Generate a pause with the minimum
    length //accoding to Table 1
     22. KH_pause =
        generate_frame(min_pause_len)
     23. seg_pause[i]+= KH_pause
24. END
```

At the step 10, to check the condition, we implement the following formula:

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

$$t = \sum_{k=1}^{M-1} (t_{1k} + t_{2k}) + t_{1M}.$$

where, t_{1k}, t_{2k} is the length of the voice and the pause calculated at step k.

IV. EXPERIMENT

We experimented the proposed algorithm with 50 different speech. The results are illustrated in Table 4. The experiment shown that our hypotheses about the minimum length can be used to segment voice for Vietnamese speech.

In general, the voice length increase when people change the speaking mode from slow, to normal and fast; and vice versa for the pause length.

Table- IV. Experiment result

	Voice length	Pause length
Slow		
1	0.331983	0.133603
2	0.340819	0.108178
3	0.376517	0.142511
4	0.305262	0.124696
5	0.338865	0.124619
6	0.316598	0.084615
7	0.314169	
...	....	..
Tmin	0,245208	0,09102
Tmax	0.350819	0.142511
Ttb	0,25781	0,11432
Normal		
1	0.330124	0.030009
2	0.350101	0.106698
3	0.313424	0.070102
4	0.333459	0.050012
5	0.366773	0.033343
6	0.300094	0.021334
7	0.316764	
...	....	....
Tmin	0,236508	0,03009
Tmax	0.367793	0.118752
Ttb	0,255634	0,080781
Fast		
1	0.308905	0.018545
2	0.306542	0.061817
3	0.303451	0.071019
4	0.307269	0.057044
5	0.305633	0.043272
6	0.306316	0.018535
7	0.296542	
...	...	....
Tmin	0,215890	0,018545
Tmax	0.306527	0.098523
Ttb	0,252091	0,060307

The result also demonstate that the value length for each speaking mode is quite the same. Therefore, we can use it for the futher task in Vietnames speech anlaysis and recognition.

V. CONLCUSION

The segmentation for Vietnamese speech is a difficult task. Because the Vietnamese speech has complex structure and intonation. In this paper, we propose an adaptive method for Vietnamese speech segmentation that based on a simple technique with an effective algorithm. We firstly estimate, by experimenting, the min and average length of a voice and a silent pause for Vietnamese speech in three main type speaking (slow, normal and fast). Then, based on these values, we start to segment the voice and pause by sliding window with proposed algorithm. The experiment result shows that the proposed method can be applied for further Vietnamese speech analysis. In the future, we will experiment the proposed method with more participants and also apply it for the speech recognition problem.

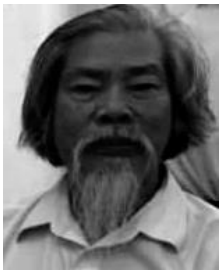
REFERENCES

1. Adami AG, Hermansky H. Segmentation of Speech for Speaker and Language Recognition OGI School of Science and Engineering , Oregon Health and Science University , Portland , USA International Computer Science Institute , Berkeley , California , USA. Eurospeech. 2003;1–4.
2. Hwa-Froelich D, Hodson BW, Edwards HT. Characteristics of Vietnamese Phonology. Am J Speech-Language Pathol. 2006;11(3):264–73.
3. Scharenborg O, Wan V, Ernestus M. Unsupervised speech segmentation : An analysis of the hypothesized phone boundaries. 2010;(February).
4. I. Mporas, T. Ganchev, and N. Fakotakis, “A hybrid architecture for automatic segmentation of speech waveforms,” in 2008 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2008, pp. 4457–4460.
5. B. Zio´lko, S. Manandhar, R. C. Wilson, and M. Zio´lko, “Wavelet method of speech segmentation,” in Signal Processing Conference, 2006 14th European. IEEE, 2006, pp. 1–5.
6. S. M. Siniscalchi, P. Schwarz, and C.-H. Lee, “High-accuracy phone recognition by combining high-performance lattice gen- eration and knowledge based rescoring,” in 2007 IEEE Interna- tional Conference on Acoustics, Speech and Signal Processing- ICASSP’07, vol. 4. IEEE.
7. Sharma A, Kaur A. Automatic Segmentation of Punjabi Speech Signal Using Group Delay. 2013;13(12).
8. Nagarajan T, Murthy HA, Hegde RM. Segmentation of speech into syllable-like units. EUROSPEECH 2003 - 8th Eur Conf Speech Commun Technol. 2003;(October 2014):2893–6.
9. Prasad, V. Kamakshi et al. “Automatic segmentation of continuous speech using minimum phase group delay functions.” Speech Communication 42 (2004): 429-446.
10. Malcangi M. Softcomputing approach to segmentation of speech in phonetic units. INTERNATIONAL JOURNAL OF COMPUTERS AND COMMUNICATIONS. 2009;3(3):41–8.
11. Md Mijanur Rahman and Md. Al-Amin Bhuiyan, “Continuous Bangla Speech Segmentation using Short-term Speech Features Extraction Approaches” International Journal of Advanced Computer Science and Applications(IJACSA), 3(11), 2012.

http://dx.doi.org/10.14569/IJACSA.2012.031121
12. T.Jayasankar et al. “Automatic Continuous Speech Segmentation to Improve Tamil Text-to-Speech Synthesis.” (2011).
13. Chowdhury, Nipa. “Separating Words from Continuous Bangla Speech T.” (2010).
14. Stolcke, Andreas et al. “Highly accurate phonetic segmentation using boundary correction models and system fusion.” 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (2014): 5552-5556.
15. Subhashini PPS, Ram MSS, Rao DS. Acoustic Feature Extraction and Optimized Neural Network Based Classification for Speaker Recognition. Int J Innov Technol Explor Eng. 2019;8(9):2496–505.
16. Sindhu B, Sujatha B. Voice Recognition System Through Machine Learning. Int J Innov Technol Explor Eng. 2019;8(10):4478–83, doi: 10.35940/ijitee.J1072.0881019.
17. J. Kamarauskas, “Automatic segmetation of phonemes using ar- tificial neural networks,” Elektronika ir Elektrotechnika, vol. 72, no. 8, pp. 39–42, 2015.
18. Amirgaliyev Y, Hahn M, Mussabayev T. The speech signal segmentation algorithm using pitch synchronous analysis. Open Comput Sci. 2017;7(1):1–8.
19. Nguyen T Van, Nguyen BQ, Phan KH, Do H Van. Development of Vietnamese Speech Synthesis System using Deep Neural Networks. J Comput Sci Cybern. 2019;34(4):349–63.
20. Ngo TL, Pham KL, Cao MS, Pham SB, Phan XH. Dialogue act segmentation for Vietnamese human-human conversational texts. Proc - 2017 9th Int Conf Knowl Syst Eng KSE 2017. 2017;2017-Janua:203–8.

AUTHORS PROFILE

First author: Tuan Anh Tran was born in 1980. He received a degree in Information Technology (IT) and a Master's degree in computer science from the Military Technical Academy in Hanoi in 2011. He is currently a lecturer. at Industrial College, Thanh Hoa, Vietnam. His areas of interest are sound processing and image processing, complex system modeling and simulation.



Second author: Dr. Mong Nguyen Huu was born in 1947. He received a degree in mathematics and mathematics (1972) and a master's degree in mathematics (1974) at the University of Minsk, Belorus, his doctorate in applied mathematics in informatics at Moscow University, Russia, in 1980. He is currently a senior lecturer at the Military Technical Academy, Hanoi, Vietnam. His areas of interest are complex audio, voice recognition, modeling and simulation systems.

Third Author: Dr. Khanh Nguyen Trong was born in 1982. He received his Bachelor degree in Information Technology (IT) at Hanoi University of Science and Technology in 2005, his Master degree in IT at the L'Institut de la Francophonie pour l'Informatique (old IFI) in 2008, and his Ph.D. in IT at the University of Paris VI , France, in 2013. He is currently a lecturer at Posts and Telecommunications Institute of Technology, Hanoi, Vietnam. He is also the member of WARM Laboratory - Sorbonne University, IRD, UMMISCO, JEAI WARM, F-93143, Bondy, France. His domains of interest are Machine Learning, Deep Learning, Distributed System, Computer Support Collaborative Work, Modelling and simulation of the complex system, Collaborative Simulation and Modelling.