HMM-based TTS for Hanoi Vietnamese: issues in design and evaluation

NGUYEN Thi Thu Trang[1, 2], *Christophe D'ALESSANDRO*[1], *Albert RILLIARD*[1], *TRAN Do Dat*[2]

[1] LIMSI-CNRS (UPR 3251), France
[2] MICA, HUST - CNRS/UMI2954 - Grenoble INP, Vietnam

trangntt@limsi.fr, cda@limsi.fr, albert.rilliard@limsi.fr, do-dat.tran@mica.edu.vn

Abstract

This paper presents the development and evaluation of an HMM-based TTS system for the modern Hanoi dialect of Northern Vietnamese, a tonal language. A study of specific phonetic and prosodic features of Hanoi Vietnamese is discussed. Consequences on the design of an HMM-based TTS system are proposed. Based on this knowledge, a TTS system, called VTed, is then developed under Mary TTS platform. The second part of the paper is devoted to perceptual evaluations of Vietnamese speech synthesis. Three kinds of evaluations are sought necessary for quality assessment of this tonal language. General MOS assessment, utterance-level intelligibility and tone-level intelligibility tests are conducted with a "natural speech reference" condition. The results show 1.21 points difference between natural and synthetic speech for MOS, from 2.5 to 4% difference for utterance-level intelligibility, and – 25% in average and depending on the tone type – from 0 to 37% difference for tone-level intelligibility. These results demonstrate the need for more specific works on the tonal/prosodic level to improve automatic synthesis of Vietnamese and other tonal languages.

Index Terms: perceptual evaluation, text-to-speech, speech synthesis, tonal language, Vietnamese, Hanoi

1 Introduction

Statistical parametric speech synthesis is most simply described as generating the average of some sets of similarly sounding speech segments. One instance of these techniques, called hidden Markov model (HMM)-based speech synthesis, has recently been demonstrated to be very effective in synthesizing acceptable speech [1]. There are some works on development of an HMM-based speech synthesis for tonal languages [2] [3] [4], but a few for Vietnamese. To the extent of our knowledge, there are two groups working on HMM-based Vietnamese speech synthesis: One from IoIT, Vietnam [5] [6] and the other is from Yunnan, China [7] [8]. They both adopted HTS framework [9] to experiment for Vietnamese speech synthesis and presented only the HTS core architecture, not the whole architecture of a Text-To-Speech system.

This paper presents an entire design of an HMM-based TTS system for Vietnamese, and our study of phonetic and phonological features of Vietnamese. Then VTed, a TTS system for Hanoi Vietnamese, is implemented following the design. The training phase and synthesis phase in VTed are done automatically regarding tone aspects.

There are a number of works on different methods and dimensions on evaluations of TTS systems. A more rigorous approach is directed towards the identification of auditory (perceptual) dimensions of speech quality, e.g. intelligibility, naturalness, loudness [10]. However, for tonal languages, to the best of our knowledge, this task is not received enough efforts in preparations and thorough discussions, especially on tone intelligibility. Evaluations in a TTS system for tonal languages are mostly conducted to test the naturalness or clearness by MOS test, with or without natural speech reference, such as [2] [11] for Chinese, [12] [13] for Vietnamese. Some works also mentioned the intelligibility but with general results and did not focus on tonal issues [5] [7]. There are some works focusing on tone correctness improvement for Thai [4] [14]. However, the experiment focused on improvement methods; the subjects were asked to decide whether the syllables had tones corresponding with the given texts or not. We believe that there is no special design to distinguish tones in the same context.

This paper aims to perceptual evaluations focusing on tonal aspects for a Vietnamese TTS system. The first evaluation is MOS test to ensure the quality in general of the system. The second one is the intelligibility test of utterances by syllable, tone and phoneme. The last one is devoted for test of tone intelligibility with some special designs for tone perceivability in the same context. These evaluations are conducted on the synthetic speech of VTed and natural speech reference.

The rest of this paper is organized as follows. Section 2 presents the development of an HMM-based TTS system for Vietnamese and the implementation of VTed for Hanoi Vietnamese. Section 3 describes our preparations and discusses experimental results of three kinds of perceptual evaluations on VTed and natural speech. The last section gives conclusions and potential works.

2 Development

2.1 System architecture

We propose the architecture of an HMM-based TTS system for Vietnamese language, illustrated in Figure 1. There are three parts in this architecture: Natural language processing (NLP) part, Training and Synthesis part.

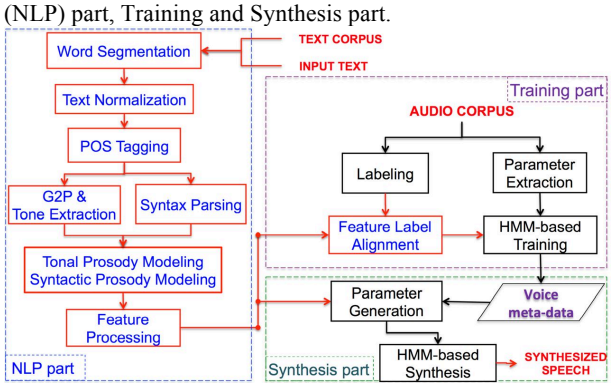

Figure 1: *Architecture of an HMM-based TTS system for Vietnamese.*

There are seven modules in the ***NLP part***, which accepts the corpus or input text and finally produces context-based features to both Training and Synthesis part. The *G2P and Tone Extraction* module produces the phonemes and tones for the text. The *Prosody Modeling* module may include both tonal and syntactic prosody modeling, based on tonal and

syntactic information from previous modules. The *Feature Processing* module accepts all information from previous modules; process them to build a set of features including contextual factors in phoneme, syllable, word, phrase and utterance level. There does exist factors related to tones such as tone type, tone features at different levels.

There are two main inputs for the ***Training part*** to produce a trained voice using HMM and EM algorithm from an audio with corresponding text corpus: (i) Speech parameters including spectral (mel-cepstrum) and excitation parameters, which are extracted from the audio corpus (ii) Context-based features from text corpus aligned with labels, which is automated labeling from the audio corpus.

In the ***Synthesis part***, context-based features are used to produce a sequence of speech parameters in such a way that its output probability for the HMM is maximized. Then using these speech parameters and synthesis filter, we can obtain high-quality synthesized speech.

2.2 Vietnamese phonetic and phonology

2.2.1 Vietnamese syllable structure

The structure of Vietnamese syllables has been the subject of strong debates. The study in [15] concluded that there are *four parts* in the structure: *initial, nucleus, ending and tone (no rhyme)*. He also argued that the *medial* should be considered as a 'semi-vowel' instead of a main part in the structure. However, [16] and [17] preferred the *hierarchical* structure and affirmed the *big role of the rhyme* in the structure of Vietnamese syllables. The authors in [16] did not take tones into account in the structure and they argued that medial (/w/) is in nucleus. For the tone, [17] assumed that the tone is a part of the whole syllable. The study in [18] did a perception test using Diagnosis Rhyme Test (DRT) method and concluded that the initial does not take part in the construction of the tone, which means that the Vietnamese tone affects only the rhyme of the syllable.

With all above analysis, we adopted the hierarchical structure with two main parts of a syllable: *An initial consonant* and *a rhyme*. A tone is one part of rhyme with three other elements: medial, nucleus and ending. The nucleus and tone are compulsory while others are optional.

2.2.2 Vietnamese phonological system

Table 1: *Vietnamese initial consonants*

Manner of articulation \ Place of articulation			Bi-labial	Labio-dental	Alveo-lar	Pala-tal	Ve-lar	Glo-tal
Nasal			m		n	ɲ	ŋ	
Plosive	**Aspirated**				tʰ			
	Un-aspirated	**Voiceless**	p		t	c	k	ʔ
		Voiced	ɓ		ɗ			
Fricative		**Voiceless**		f	s		x	
		Voiced		v	z		ɣ	h
Approximant (Liquid)					l			

Table 1 presents our conclusion for Hanoi Vietnamese initial consonants. In this dialect, orthographic ch /c/ and tr /ʈ/, d /z/ and gi, r /ʐ/, x /s/ and s /ʂ/ are pronounced alike as /c/, /z/, /s/ [19] [20]. Hanoi Vietnamese licenses eight segments in coda position: three unreleased voiceless obstruents /p t k/, three nasals /m n ŋ/, and two approximants /j w/ (semi-vowels). Following back rounded vowels /u o ɔ/, the velar stops /k N/ are produced as doubly articulated labial-velars /kp ŋm/ [21].

Hanoi Vietnamese distinguishes nine short vowels /i e ɛ a ɯ ɤ u o ɔ/, four short vowels /ɛ̆ ă ɤ̆ ɔ̆/ and three falling diphthongs /ie ɯɤ uo/ [17], illustrated in Table 2.

Table 2: *Vietnamese vowels/diphthongs*

Elevation \ Position	Front	Central	Back	
			Unrounded	Rounded
Close (High vowel)	i ie		ɯ ɯɤ	u uo
Close-mid	e		ɤ ɤ̆	o
Open-mid	ɛ ɛ̆			ɔ ɔ̆
Open (Low vowel)		a ă		

2.2.3 Vietnamese tones

Northern Vietnamese is known as a tonal language having six different lexical tones. These are: Level (1), falling (2), broken (3), curve (4), rising (5), and drop (6) tones. The study of [22] [23] confirms that voice quality is a robust correlate of tone in Hanoi Vietnamese, showing less variability than F0 across reading conditions. The experiment in [23] warrants the conclusion that rising (5b) and drop (6b) tones of syllables ending in /p/, /t/ or /k/ (checked syllables) are not glottalized, either in final or non-final position. The work on oral flow [24] brings out a clear difference between these two sets of rhymes: tone 6a (drop tone in unchecked syllables) has low oral airflow; tone 5b and 6b have relatively high oral airflow, getting close to the range of breathy voice.

Table 3: *Characteristics of 8 Vietnamese tones*

No.	Name	Register	Duration	F0 contour	Phonation
1	***Level***	High (+)	Long (+)	Level	Modal voice
2	***Falling***	Low (-)	Long (+)	Falling	Breathy voice
3	***Broken***	High (+)	Long (+)	Falling-Rising	Glotallization
4	***Curve***	Low (-)	Long (+)	Falling	Harsh voice
5a	***Rising***	High (+)	Long (+)	Rising	Modal voice
5b	***Rising***	High (+)	Short (-)	Rising	Tense voice
6a	***Drop***	Low (-)	Short (-)	Dropping	Glottalization
6b	***Drop***	Low (-)	Short (-)	Dropping	Tense voice

Therefore, it could be said that Vietnamese has a six-tone paradigm for sonorant-final syllables, and a two-tone paradigm for obstruent-final syllables [23], summarized in Table 3.

2.2.4 Vietnamese allophones

As presented in subsection 2.2.2, there are totally 21 consonants, 13 vowels, three diphthongs and two semi-vowels in Hanoi Vietnamese. However, if we also consider the function of each phoneme in the syllable, there are 19 allophones for initial, one allophone for medial, 18 allophones for nucleus and 10 allophones for ending (including two semi-vowels).

Table 4: *Methods to build the allophone set for Vietnamese*

Function of phone?	Embed tone?	Results	Allophones #
No	No	21 consonants, 13 vowels, 3 diphthongs & 2 semi-vowels	39
Yes	No	19 initials, 1 medials, 18 nucleus and 10 endings	48
Yes	Yes, in nucleus	19 initials, 1 medials, 18x6 nucleus and 10 endings	138
Yes	**Yes, in all elements of the rhyme**	**19 initials, 1x6 medials, 18x6 nucleus and 10x6 endings**	**193**

In fact, based on the discussion of syllable structure, the tone is produced in parallel with each phoneme in the rhyme

during the articulation. It could be said that the same phoneme in the same rhyme bearing different tones is unalike. HMM-based speech synthesis models parameter sequences, sequences of allophones - not in parallel. Therefore, beside tonal factors, it is wiser to embed tones inside suitable phonemes for syllables. Table 4 summarizes different ways to take tones into account to build up the allophone set. After doing a preliminary evaluation for each method, we adopted the last method, which produces the best quality for tone synthesis.

2.3 Implementation of VTed

There are several platforms that can be used to develop an HMM-based TTS system [9]. We have chosen Mary TTS platform, because of their facilities and ease of expandability, to build an HMM-based TTS system for Hanoi Vietnamese, VTed.

2.3.1 Implementation of VTed

We have built a NLP part integrating with other modules in Mary TTS to establish an automatically training and synthesis process. All modules are implemented following the architecture in Figure 1. We adopted the method in [25] to build the *Word Segmentation*, [26] to build the *POS Tagging* module. For the *Text Normalization* module, we integrated our previous work [27] into the NLP part. The *Prosody Modeling* module is simple applied the ToBI model to get features related to ToBI endtone of syllables and phrases. The *G2P & Tone Extraction* module is designed specially to easily change different allophone units. The *Feature Processing* module and *Feature Label Alignment* module are modified from the existed modules in Mary TTS so that it is suitable for Vietnamese.

Other modules are reused from Mary TTS platform as the following main points: (a) The *Labeling* module: Using eHMM tool of Festvox to label audio corpus automatically. (b) The *Parameter Extraction* module: Using the SPTK and Snack tools for extracting features as in the original HTS scripts. (c) The *HMM-based Training* module: Using a version of the speaker dependent training scripts provided by HTS that (i) use context features predicted by the NLP part, (ii) include global variance (iii) compose training data from mel-generalised cepstrum (mgc), log F0 and strength files and (iv) include features (band pass voicing strengths) for generation of mixed excitation [28]. (d) The *Parameter Generation* module and *HMM-based Synthesis* module: a new HMM-based synthesizer ported to Java from the HTS [29].

2.3.2 VN-Voice Training

There are 84.31% of words composing of at least two syllables [25]. Based on the syllable structure in subsection 2.2.1, there are four elements at phoneme-level. Therefore, we used 5-state left-to-right HMMs with single diagonal Gaussian output distributions to capture all the elements of syllables and boundaries of syllables and words.

The training is automatically carried out with a corpus of ~92% from 630 sentences from our existed corpus, VNSpeechCorpus, while the rest (8%) are used in the evaluation phase. These sentences are recorded by a Vietnamese female broadcaster from Hanoi at 48 kHz and 16 bits per sample. Total duration of all sentences is about 37 minutes.

3 Perceptual evaluation

Our perceptual evaluations include the assessment of general MOS, intelligibility of utterances and its elements and tone perceivability in context. These evaluations are carried out with VTed and natural speech reference in random order. All participated subjects are from the North of Vietnam, living for a long time in Hanoi. Participants were 20-35 years old and reported normal hearing and vision.

3.1 Evaluation of general quality

Subjects were asked to give their scores *"5-Excellent, 4-Good, 3-Fair, 2-Poor and 1-Bad"* for overall impression after listening to an utterance. There are 48 sentences in test corpus (8% of VNSpeechCorpus). To have more reference, this test is also carried out on our previous TTS system adopting non-uniformed unit-selection synthesis - HoaSung [13] with the same corpus (92% of VNSpeechCorpus).

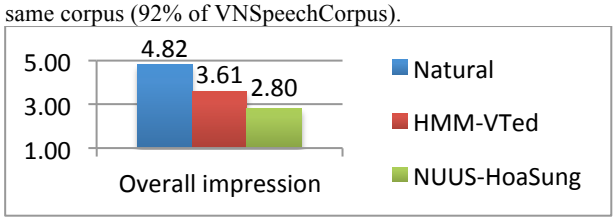

Figure 2: *Results of quality in general (MOS Test).*

A two-factorial ANOVA was run on the results. The 2 factors were the TTS system (3 levels) and the Sentence (48 levels). All factors (and their interaction) have highly significant effect ($p<0.001$); meanwhile the TTS system factor alone explains about 63% of the variance (partial $\eta^2=0.63$), while the Sentence factor and the interaction explain only about 15% each. A post-hoc Tukey test shows that each TTS system received significantly different mean scores. The experiment results plotted in Figure 2 shows the sound quality of VTed is rather good (0.81 point higher than HoaSung), but still clearly distinguishable 1.21 point lower) from natural speech.

3.2 Evaluation of utterance-level intelligibility

Subjects were asked to write down texts of utterances they heard. They can listen to one, two or three times. To prevent subjects from listening the same text from different systems, we adopt a Latin square design (Cochran and Cox 1992). We prepare 74 sentences with a length of 8 to 20 syllables, found in newspapers. 30 subjects took the test (15 females). Each subject is presented with half natural utterances and half VTed sentences.

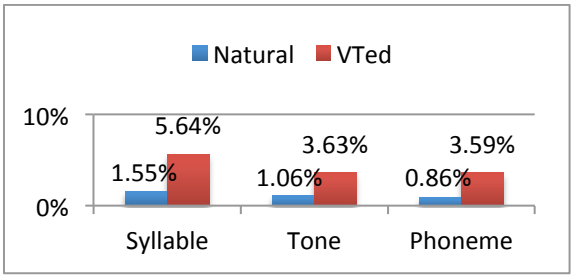

Figure 3: *Error rates of intelligibility in utterance elements.*

To measure the error rate, we adopted a metric based on *approximate* string matching [30] and the Damerau-Levenshtein algorithm. We did some improvements in

dynamic programming to calculate the distance between two strings, and then error rates based on the metric proposed in [31] for syllable, tone and phoneme.

The error rates of VTed and natural speech are illustrated in Figure 3. VTed diverges from natural speech from 2.6-4.1%.

3.3 Evaluation of tone intelligibility

There are 18 subjects (9 females and 9 males) participating the test of tone intelligibility. Subjects can decide to listen to one more time or not after their selection for the first listening time.

3.3.1 Stimuli and paradigm

In this evaluation, groups of meaningful sentences with the same syllables, diverging only for one tone, are prepared. As a systematic variation of tones on the same syllable is rare, so we need to use proper names for some special cases. Subjects need to choose the most likely syllable they heard among a group of syllables bearing different tones in an utterance. E.g.

- Ở đây có buôn bán … không? *(Do you sell … here?)*
  - dê *(goats – level tone 1)*
  - dễ *(easily – broken tone 3)*
  - dế *(crickets – rising tone 5a)*
- Mỗi tối, bác sĩ … thường đến hỏi thăm các bệnh nhân *(Every evening, doctor … usually visits her patients)*
  - Thuỷ *(a person name with curve tone 4)*
  - Thuỳ *(a person name with falling tone 2)*
  - Thuý *(a person name with rising tone 5a)*
  - Thuỵ *(a person name with drop tone 6a)*

For each utterance, we also put the "Not like all above options" in the answer list. We have designed intently 129 sentences in 40 groups. The balance of tone pair is also calculated, which is from 9-12 examples for each tone pair.

3.3.2 Experiment results

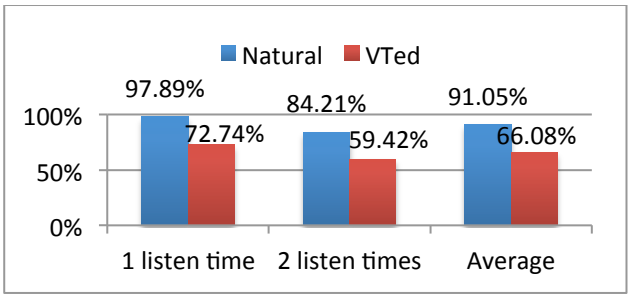

Figure 4: *Corrected rates of tone intelligibility.*

The corrected rates of tone intelligibility are illustrated in Figure 4. The results are 10% higher for sentences needing one listening time than those needing two listening times. The synthetic speech received about 25% corrected rate lower than the natural speech. The proportion of "Not like all above options" selection is really minor, from 0.0 to 0.7% for natural speech, and from 1.0 to 3.4% for VTed.

The corrected rates by tone types in Figure 5 show that all tones are nearly 100% perceivable in the same context, except the falling tone 2. We believe that this tone is the most difficult tone for both systems.

The result of VTed can be divided into three groups of tones. The first group, 100% perceivable as natural speech, is for the rising tone in checked syllable. The second one, from 18 to 22% lower than natural speech, is for the level tone 1, drop tone in checked syllable 6b (around 87% perceivable) and the rising and drop tone in unchecked syllable 5a, 5b (around 80% perceivable). The last group, from 29 to 37% lower than natural speech, is for the broken tone 3, curve tone 4 and falling tone 2 (around 62% perceivable).

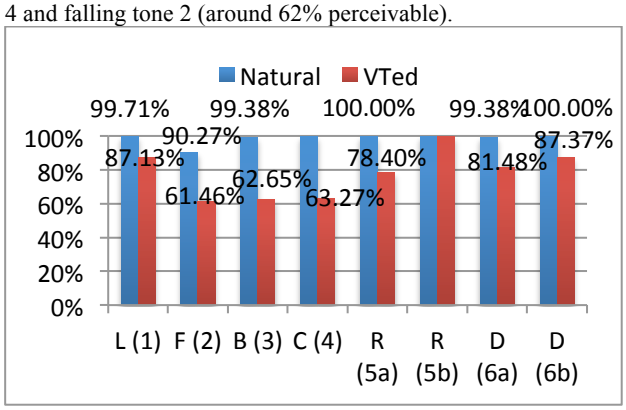

Figure 5: *Corrected rates by tone types of tone intelligibility.*

4 Conclusions and perspectives

Based on our study on Vietnamese phonetic and phonological features and design of a TTS system for Vietnamese, an HMM-based TTS system for Hanoi Vietnamese (VTed) is developed. The allophone set is built at phoneme-level regarding the tone type and the function of phoneme in rhyme. Context-based features, including tone features, are extracted automatically for both training and synthesis phase in VTed. Three kinds of tests are conducted to assess the quality of VTed. In the MOS Test, the score of VTed is 1.21 points lower than natural speech, while the error rate of utterance-level intelligibility is only from 2.5-4% higher than the natural speech. These results show that VTed can produce a rather good speech. In the tone intelligibility, subjects need to perceive different tones in the same context. The corrected rate of VTed in this test is 25% lower than natural speech in average. Participators can perceive nearly 100% the rising tone in checked syllables for both VTed and natural speech. For other tones, the differences between VTed and natural speech are from 18-37%.

These results show the need for more works to improve the quality of an HMM-based Vietnamese TTS system. Some studies on corpus design for Vietnamese HMM-based speech synthesis will be done in the future. This design should consider the allophone unit (e.g. initial consonants and rhyme instead of phoneme-level) and the balance of allophones/tones, allophones in (tonal) context. Moreover, we should spend efforts on tonal and/or syntactic prosody modeling in NLP part, which provides more specific context-based features of Vietnamese for both training and synthesis processes, instead of ToBI model. Due to the phenomena of glottalization in Vietnamese, this task may study the influence of not only F0 evolution, but also intensity and duration on tones in context. In addition, tone correctness can be improved by design of decision-tree structure based on tone features, as the works for Thai [4]; or directly interference of input excitation parameters for training and synthesis process.

5 Acknowledgements

We would like to thank Lê Quang Thắng and Nguyễn Thị Lan for their supports in recording and conducting tests at MICA.

6 References

[1] H. Zen, K. Tokuda, and A. W. Black, "Statistical parametric speech synthesis," *Speech Communication*, vol. 51, no. 11, pp. 1039–1064, 2009.

[2] Y. Guan, J. Tian, Y. Wu, J. Yamagishi, and J. Nurminen, "An Unified and Automatic Approach of Mandarin HTS System," in *Proceedings of the 7th Speech Synthesis Workshop (SSW7)*, Kyoto, Japan, 2010.

[3] S. P. Ya Li, "HMM-based speech synthesis with a flexible Mandarin stress adaptation model," in *Proceedings of the IEEE 10th International Conference on Signal Processing (ICSP)*, Beijing, China, 2010, pp. 625 – 628.

[4] S. Chomphan and T. Kobayashi, "Tone correctness improvement in speaker dependent HMM-based Thai speech synthesis," *Speech Communication*, vol. 50, no. 5, pp. 392–404, 2008.

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

[6] T. S. Phan, T. T. Vu, T. C. Duong, and C. M. Luong, "A study in Vietnamese statistical parametric speech synthesis base on HMM," *International Journal of Advances in Computer Science and Technology*, vol. 2, pp. 1–6, 2012.

[7] L. Kui, J. Yang, B. He, and E. Hu, "An Experimental Study on Vietnamese Speech Synthesis," in *Proceedings of the International Conference on Asian Language Processing (IALP)*, Penang, Malaysia, 2011, pp. 232–235.

[8] L. He, J. Yang, L. Zuo, and L. Kui, "A trainable Vietnamese speech synthesis system based on HMM," in *Proceedings of the International Conference on Electric Information and Control Engineering (ICEICE)*, Wuhan, China, 2011, pp. 3910–3913.

[9] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," presented at the Proceedings of the 6th ISCA Workshop on Speech Synthesis, Bonn, Germany, 2007, pp. 294–299.

[10] U. Jekosch, "Speech quality assessment and evaluation," in *Proceedings of the European Conference on Speech Communication and Technology*, Berlin, Germany, 1993.

[11] M. Dong, K. Lua, and H. Li, "A Unit Selection-based Speech Synthesis Approach for Mandarin Chinese," *Journal of Chinese Language and Computing*, vol. 16, pp. 135–144, 2006.

[12] D. D. Tran and E. Castelli, "Generation of F0 contours for Vietnamese speech synthesis," in *Proceedings of the third International Conference on Communications and Electronics (ICCE)*, Nha Trang, Vietnam, 2010, pp. 158–162.

[13] V. T. Do, D. D. Tran, and T. T. T. Nguyen, "Non-uniform unit selection in Vietnamese speech synthesis," in *Proceedings of the Second Symposium on Information and Communication Technology*, Hanoi, Vietnam, 2011, pp. 165–171.

[14] S. Chomphan and C. Chompunth, "Improvement of Tone Intelligibility for Average-Voice-Based Thai Speech Synthesis," *American Journal of Applied Sciences*, vol. 9, no. 3, pp. 358–364, 2012.

[15] X. K. Đoàn, "Xem lại một vấn đề ngữ âm tiếng Việt: Cấu trúc âm tiết (Re-consider a problem of Vietnamese phonetics: Syllable structure)," *Hợp Lưu*, vol. 48, pp. 1–24, 1999.

[16] I. Vogel, I.-J. E. Tseng, and N.-T. Yap, "Syllable structure in Vietnamese," in *Proceedings of the second Theoretical East Asian Linguistic (TEAL) Workshop*, Taiwan, 2004.

[17] Đoàn T. T., *Ngữ Âm Tiếng Việt (Vietnamese phonetics)*. Đại Học và Trung Học Chuyên Nghiệp, 1999.

[18] D. D. Tran, E. Castelli, J.-F. Serignat, V. L. Trinh, and X. H. Le, "Influence of F0 on Vietnamese syllable perception," in *Proceedings of the 9th European Conference on Speech Communication and Technology*, Lisbon, Portugal, 2005, pp. 1697–1700.

[19] L. C. Thompson, *A Vietnamese Reference Grammar*. University of Hawaii Press, 1987.

[20] A.-G. Haudricourt, "'The origin of the peculiarities of the Vietnamese alphabet' (Translated by Alexis Michaud)," *Mon-Khmer Studies*, vol. 39, pp. 89–104, 2010.

[21] J. P. Kirby, "Vietnamese (Hanoi Vietnamese)," *Journal of the International Phonetic Association*, vol. 41, no. 03, pp. 381–392, 2011.

[22] V. L. Nguyễn and J. A. Edmondson, "Tones and voice quality in modern northern Vietnamese: Instrumental case studies," *Mon-Khmer Studies Journal*, vol. 28, pp. 1–18, 1998.

[23] A. Michaud, "Final consonants and glottalization: new perspectives from Hanoi Vietnamese," *Phonetica*, vol. 61, no. 2–3, pp. 119–146, 2004.

[24] 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 Studies*, vol. 36, pp. pp. 121–137, 2006.

[25] H. P. Le, T. M. H. Nguyen, A. Roussanaly, and T. V. Ho, *A Hybrid Approach to Word Segmentation of Vietnamese Texts*, vol. 5196. Springer-Verlag Berlin, Heidelberg ©2008, 2008.

[26] H. P. Le, A. Roussanaly, T. M. H. Nguyen, and M. Rossignol, "An empirical study of maximum entropy approach for part-of-speech tagging of Vietnamese texts," in *Traitement Automatique des Langues Naturelles - TALN 2010*, Montreal, Canada, 2010.

[27] T. T. T. Nguyen, T. T. Pham, and D. D. Tran, "A method for Vietnamese text normalization to improve the quality of speech synthesis," in *Proceedings of the 2010 Symposium on Information and Communication Technology*, Hanoi, Vietnam, 2010, pp. 78–85.

[28] S. Pammi, M. Charfuelan, and M. Schröder, "Multilingual Voice Creation Toolkit for the MARY TTS Platform," in *Language Resources and Evaluation (LREC)*, Malta, 2010.

[29] M. Schröder, M. Charfuelan, S. Pammi, and O. Türk, "The MARY TTS entry in the Blizzard Challenge 2008," in *Blizzard Challenge 2008*, Queensland, Australia, 2008.

[30] G. Navarro, "A Guided Tour to Approximate String Matching," *ACM Computing Surveys*, vol. 33, pp. 31–88, 2001.

[31] R. W. Soukoreff and I. S. MacKenzie, "Measuring errors in text entry tasks: an application of the Levenshtein string distance statistic," in *Proceedings of the CHI '01 Extended Abstracts on Human Factors in Computing Systems*, New York, USA, 2001, pp. 319–320.