Prosodic phrasing modeling for Vietnamese TTS using syntactic information

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

[1] LIMSI-CNRS (UPR 3251), France
[2] MICA Institute, HUST - CNRS/UMI2954 - Grenoble INP, Hanoi, Vietnam
trangntt@soict.hust.edu.vn, albert.rilliard@limsi.fr, do-dat.tran@mica.edu.vn, cda@limsi.fr

10.21437/Interspeech.2014-192

Abstract

This research aims at modeling prosodic phrasing for improving the naturalness of Vietnamese (a tonal language) speech synthesis. The proposed phrasing model includes hypotheses on: (i) prosodic structure based on syntactic rules (ii) final lengthening linked to syllabic structures and tone types. Audio files in the analysis corpus are manually transcribed at the syllable level and perceived pauses. Text files are parsed and represented with annotated-syntax trees. Statistical treatment brings out a correlation between syntactic element boundaries and pause duration. Major breaks may appear at the end of a clause or between predicates or head elements. Other rules between grammatical phrases/words or shorter clauses may trigger minor breaks. Break levels (including ones predicted by syntactic rules) and relative positions of syllables are used to train VTed, an HMM-based Text-To-Speech (TTS) system for Vietnamese. In the synthesis phase, break levels are explicitly inserted while lengthening is applied for last syllables of prosodic phrases. Perceptive testing shows an increase of 0.34 on a 5 point MOS scale, for the new prosodic informed system (3.95/5) compared to the previous TTS system (3.61/5). In the pair-wise comparison test, about 70% of the synthetic voice with the proposed model is preferred to the previous version.

Index Terms: prosody modeling, phrasing, final lengthening, text-to-speech, speech synthesis, tonal language, Vietnamese

1 Introduction

Phrasing modeling plays an important role in improving the naturalness for speech synthesis. Many researchers have been working on prosodic structure generation for Chinese [1][2], pause/break modeling for French [3], German [4], Russian [5] or modeling style specific break [6][7]. They may use rules or machine learning with lexical information (e.g. POS tagger) or contextual length. However, to the best of our knowledge, there is no such work for Vietnamese, a tonal language. It is believed that there is an interface between syntax and prosodic structure [8][9][10][11][12]. Recently, much effort has been devoted for Vietnamese syntax parsing with some has proven fruitful results [13][14][15][16]. This paper aims at modeling prosodic phrasing using syntactic information for Vietnamese speech synthesis. In this work, final lengthening of phrases is also considered since it is a crucial aspect of the naturalness of areas around boundaries of speech [17][18].

Audio files in the analysis corpus are manually transcribed, time-aligned at the syllable level, and annotated for perceived pauses. Text files are parsed and represented with annotated-syntax trees. Statistical analyses are carried out to find a correlation between syntactic element boundaries and pause duration as well as final lengthening. Durations of next pauses and last syllables of predicted phrases are measured. Final lengthening of last syllables is calculated based on z-score normalization, linked to syllabic structures and tone types. The proposed model is implemented in VTed, an HMM-based TTS system for Vietnamese [19]. Break levels (including ones predicted by syntactic rules) and relative positions of syllables are used to train VTed with other prosodic features. In the synthesis phase, predicted boundaries with relevant break levels are explicitly inserted whereas lengthening is applied for last syllables of predicted phrases. Two types of perception test are conducted: (i) MOS test for naturalness using a natural speech reference, (ii) Pair-wise comparison test for evaluating the preference of the new prosodic informed system to the previous one. Some statistical analyses are carried out the see the difference in the results.

The rest of this paper is organized as follows. Section 2 presents the background of this work. Section 3 presents our proposal on prosodic phrasing using syntactic information with some analysis results. The implementation and evaluation of the proposed model are presented in the Section 4, 5. The final section gives conclusions and presents future works.

2 Background

2.1 Vietnamese phonetics and syntax

We adopted a hierarchical structure for Vietnamese syllables, based on an initial consonant (C) and a rhyme. Tone is carried by the rhyme on 3 elements: medial (w), nucleus (V) and ending (C). Nucleus and tone are compulsory while others are optional. Vietnamese has a six-tone paradigm (level 1, falling 2, broken 3, curve 4, rising 5a, and drop 6a) for sonorant-final syllables, and a two-tone paradigm (rising 5b, drop 6b) for obstruent-final ones. For duration of bearing syllables, there are 2 kinds of tones: (i) long tones: 1-4, 5a, and (ii) short tones: 5b, 6a, 6b. Details are presented in [19][20].

We adopted the part-of-speech (POS) tag set from the Vietnamese POS tagger [21] and phrasal categories from the VietTreeBank, a Vietnamese syntactically-annotated corpus [14]. Major POS categories are noun (N), proper noun (Np), verb (V), adjective (A) and preposition (P) whereas minor ones are conjunction (CJ), interjection (I). A phrase includes one or more heads (phrase head – H, giving name to the phrase), proceeding and succeeding dependent constituents. For instance, the noun "người" *(man)*, which determines the phrase name (noun phrase – NP), is the phrase head of "một người cao lớn" *(a tall man)*. Some other main phrasal categories are PP (preposition phrase), VP (verb phrase) and AP (adverb phrase). In the syntactic structure of sentences, two distinct yet interrelated aspects must be distinguished [22] [23]: (i) Phrase structure grammar concerns the organization of the units that constitute sentences, e.g. S → PP + NP + VP (ii) Dependency grammar encompasses the dependency relation, e.g. subject–predicate. Labels for both types of grammar are adopted from the VietTreeBank, with adaptations.

In a coordinate sentence, two or more main clauses (S) occur as daughters and co-heads of a higher clause. A subordinate clause (SBAR, e.g. relative clause) is one that functions as a dependent, rather than a co-head [23]. The set of

types (i.e. sentence, clause, phrase, word, and morpheme) is structurally organized in a part–whole hierarchy: each unit is entirely composed of smaller units belonging to a limited set of types. The following example illustrates this hierarchical structure: [$_S$ [$_{NP}$ [$_N$ Cô giáo *(The teacher)*] [$_{Np}$ tiếng Anh *(English)*] [$_{SBAR}$mà *(who)* [$_{NP}$ [$_N$ anh *(you)*]] [$_{VP}$ đã [$_V$ gặp *(met)*]] [$_{NP}$ [$_N$ hôm qua *(yesterday)*]] $_{SBAR}$] $_{NP}$] [$_{VP}$ đang [$_V$ đọc] (*is reading*) [$_{NP}$ [$_N$ sách *(books)*]] [$_{PP}$ [$_P$ trong *(in)*] [$_{NP}$ [$_N$ thư viện *(the library)*]$_{NP}$]$_{PP}$] $_{VP}$] $_S$].

The elements of a simple clause (aside from the predicate – PRD itself) can be classified as either adjuncts (ADT) or arguments. Adjuncts are elements that are optional and not closely related to the meaning of the predicate but which are important to help the hearer understand the flow of the story, the time or place of an event, etc. Arguments are those elements that are required or allowed by some predicates, but not by others. In order to be expressed grammatically, arguments must be assigned a grammatical relation within the clause. There are two basic classes of grammatical relations: obliques (or indirect arguments) vs. terms (or direct arguments). Terms (i.e. subject – SUB, direct object – DOB, indirect object – IOB) play an active role in a wide variety of syntactic constructions, while obliques (OBL) are relatively inert. Some dependent elements are illustrated in the following example: [$_S$[$_{ADT}$ Tối qua *(last night)*] [$_{SUB}$ Kiên *(Kien)*] [$_{PRD}$ đã tặng *(gave)* [$_{DOB}$ một bó hoa hồng *(a bouquet of roses)*] [$_{OBL}$ cho mẹ của anh ấy *(to his mother)*] $_{PRD}$] $_S$].

2.2 Prosodic hierarchy

A crucial problem has been to develop a theory of syntactic juncture that can predict the domains in which rules are bound, and locate the points in syntactic structure that trigger the phonological rules. One particularly interesting theory for this sort is "prosodic hierarchy", proposed in [8][9], and extended by [10][24], together with the development of metrical theory [11], summarized in [12]. We assume here a theory involving five levels of hierarchical structure (i.e. upper levels can include one or more lower levels): the Utterance (U), the Intonation Phrase (IP), the Phonological Phrase (PhP) and the Word (W) [11], illustrated in Table 1. U and W boundaries can be identified automatically by sentence punctuations (e.g. ". ? !") or by edges of grammatical words, respectively. There is no explicit rule for the Vietnamese PhP, or IP boundaries. In the next section, we present our hypotheses and analyses results for these boundaries (hereafter called intermediate boundaries) using syntactic information.

Table 1: *Interface of prosody hierarchy and syntax*

Prosodic structure	Correlation to syntax	
Utterance boundary (U)	full sentence, comprising maximal sequence between structural pauses	
Intonation Phrase (IP)	e.g. boundaries of clauses	***=> Varying in their application and hard to pin down***
Phonological Phrase (PhP)	e.g. boundaries of phrases	
Word (W)	Grammatical word	

3 Phrasing modeling

3.1 Corpus preparation

Vietnamese is an under-resourced language, especially for speech processing [25]. For a preliminary experiment on prosodic phrasing modeling, we adopted the existing corpus, "VNSpeechCorpus for speech synthesis" [26]. This corpus contains 630 sentences (~37 minutes) recorded by a female broadcaster from Hanoi at 48 kHz and 16bps. Audio files are transcribed, time-aligned at the syllable level, and annotated for perceived pauses. Text files are parsed and represented with syntax trees in XML (eXtensible Markup Language) format. These tasks were semi-automatically executed.

3.2 Proposed syntactic rules

Two types of rules were used: one between two constituents in phrase structure grammar and the other between two elements in dependency grammar. Proposed rules for syntactic constituency and for syntactic dependency are presented respectively in Table 2 and Table 3. IP boundaries are set if either the left constituent is or contains a clause (HC1, HC2) or both left and right dependent elements are predicates (HD1) or head elements (HD2). Other decisions are made on the basis of syntactic information or number of syllables in the left or right elements. We proposed two level of Phonological Phrase (PhP1 and PhP2) due to the number of syllables in elements.

Table 2: *Constituency rules and intermediate boundaries*

#	Intermediate boundaries	Left constituent	Right constituent
HC1	*IP*	a SBAR or a constituent whose child is a clause	any constituent
HC2		a S >= 6 syllables	any constituent
HC3	*PhP level 1 (PhP1)*	a phrase >= 7 syllables	a phrase >=4 syllables
HC4		a phrase	a CJ following by a constituent >= 5 syllables
HC5		a PP >= 3 syllables	a CJ or AP/NP/VP
HC6		a S having 3 to 5 syllables or CJ following by S	a constituent
HC7		a CJ 'rằng' *('that')* whose parent is a SBAR	a constituent

3.3 Results analysis

To validate the ability of these syntactic rules to predict intermediate boundaries, proposed syntactic rules are automatically put into syntax trees to predict boundaries. All predicted boundaries are then put into TextGrid files in a tier named by the rule. Durations of last syllables preceding predicted boundaries and of pauses succeeding them are measured. Pause durations are computed in a logarithmic scale, which is more relevant to perception. Final lengthening is calculated using Z-score normalization, based on syllable structures and tone types (hereafter called syllable categories) of the last syllables.

Table 3: *Dependency rules and intermediate boundaries*

#	Boundary	Left dependent element	Right dependent element
HD1	*IP*	a PRD	a PRD
HD2		a H >= 4 syllables	a H
HD3		an ADT >= 3 syllables	any dependent element that is a phrase
HD4	*PhP level 2 (PhP2)*	a H: 2-3 syllables	a H
HD5		an ADT having 2-3 syllables	a SUB >= 2 syllables

Analyses of variance were run on pauses lengths and final lengthening. An α level of 0.05 was adopted. The fixed factors considered in each ANOVA are "Syntactic Rule" (12 levels) and "Intermediate Boundary" (3 levels). To eliminate the side effect of taking the logarithm of cases where pauses have a duration of zero (no pause), all zero pause cases (40/547) are removed from the analyses based on Log(Pause). Table 4

shows the ANOVA results. All analyses are significant, except the effect of Intermediate Boundary on Lengthening. The results show that the proposed syntactic rules to set up a hierarchy of boundaries are mainly related to the pause length, which is illustrated in Figure 1.

Table 4: *Anova results of the analysis corpus*

Anova	df	df error	F	p	η^2
Pause ~ Syntactic Rule	11	531	8.2	0.000	0.14
Log(Pause) ~ Syntactic Rule	11	486	8.3	0.000	0.16
Lengthening ~ Syntactic Rule	11	531	3.9	0.000	0.07
Pause ~ Intermediate Boundary	2	540	34.7	0.000	0.11
Log(Pause) ~ Intermediate Boundary	2	495	34.3	0.000	0.12
Lengthening ~ Intermediate Boundary	2	540	2.7	0.067	0.01

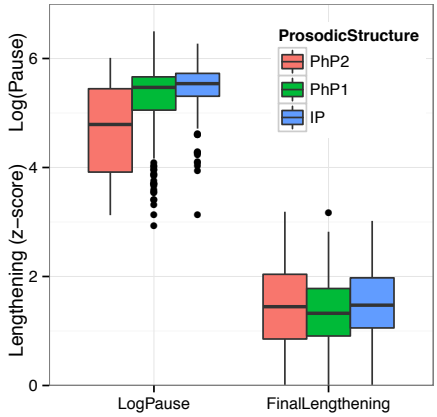

Figure 1: *Analysis results for Lengthening (bottom-right) and Log(Pause) (top-left) of Phonological Phrase level 1&2 (PhP1, PhP2) and Intonational Phrase (IP) boundaries.*

Table 5: *Detailed results for (a) rules and (b) boundaries*

(a) Syntactic rules				*(b) Intermediate boundaries*		
Rule	Percentage of non-zero pauses	Pause Mean	Group (Log (Pause))	Group (Log (Pause))	Pause Mean	Boundary (Break level)
HD3	98% (59/60)	253.24	a	a	251.75	IP (4)
HC2	95% (75/79)	250.70	a			
HD2	100% (28/28)	266.89	ab			
HC1	96% (26/27)	254.49	ab			
HD1	100% (23/23)	248.25	ab			
HC3	91% (62/68)	240.92	ab	b	218.87	PhP1 (3)
HC4	89% (48/54)	224.68	ab			
HC5	88% (45/51)	203.56	ab			
HC7	100% (08/08)	178.45	abc			
HC6	83% (29/35)	160.51	abc			
HD4	89% (79/89)	147.06	bc	c	138.52	PhP2 (2)
HD5	78% (18/23)	136.31	c			

Post-hoc Tukey tests are conducted to discover which levels of each factor show significant differences. Table 5 shows the mean values of pause length (in ms) for each rule, and the groups obtained on the basis of (a) syntactic rules or (b) intermediate boundaries, using logarithm values of pause lengths. The percentage of non-zero pauses is calculated as the ratio of non-zero pauses over the total number of boundaries that each syntactic rule finds. Syntactic rules for the IP boundaries have a high ratio of pauses (>=95%), which gradually falls down with the boundary level (except for HC7 which has a small number of observations). The grouping of syntactic rules to intermediate boundaries is coherent.

4 TTS system

4.1 Design of Vietnamese training features

VTed is an HMM-based TTS system [27] for Vietnamese, based on 3 main parts: Natural language processing (including Prosody Modeling), HMM Training, and Synthesis (cf. [19] for details). Contextual features for Vietnamese are chosen at phoneme, syllable, word, phrase, and utterance levels (for detail on these features, cf. [20]).

Based on the proposed model of prosodic phrasing, two new prosodic features were introduced: break levels and syllable position relative to phrase. Table 6 presents our proposed break levels and syllable's relative positions used as training features for the HMM-based TTS. The break levels "0", "1", "5" and "6" are easily identified by POS tags or punctuation marks at the end of sentence whereas tags "2", "3", "4" need syntactic rules for prediction. Syllable positions are distinguished for the boundaries above the Word (1) and above the Intonation phrase boundaries ("2").

Table 6: *Break levels as training features*

Break level	Syllable position	Prosodic hierarchy	Rule
0	0	Within word	Between 2 consecutive phonemes in one word
1	0	Word	Between 2 consecutive words
2	1	Phonological phrase level 2	HD4, HD5
3	1	Phonological phrase level 1	HC3, HC4, HC5, HC6, HC7
4	1	Intonation phrase	After a punctuation mark in the middle of the sentence or HC1, HC2, HD1, HD2, HD3
5	2	Utterance boundary	After punctuation marks at end of sentence, not of paragraph
6	2	Paragraph boundary	At the end of paragraph

4.2 System implementation

The training phase of VTed is automatically carried out on the VNSpeechCorpus. We use all the training features presented here, including break levels and syllable positions. In the synthesis phase, break levels are explicitly inserted using rules in Table 6. In the synthetic speech, in a preliminary test, we find that pause lengths after predicted boundaries are well modeled but final lengthening (which makes predicted boundaries disrupted and unnatural). Thus a final lengthening is then applied to the last syllables of predicted prosodic phrases. Due to the non-significant effect of lengthening on syntactic rules (i.e. predicted boundaries), we assume a lengthening amount with an average value of 135%. However, we observed that durations of some final syllables are over or under-lengthened, depending on syllable categories. Detailed information on this issue and the demonstration of the TTS system is to be found on the VTed web page [28].

For the evaluations, we prepared two versions of VTed: (i) *ProposedModel*: VTed trained with two new training features (break levels are directly predicted from syntax trees and final lengthening is estimated from syllable categories); (ii) *PreviousVersion*: the previous version of VTed [19][20].

5 Evaluation

The perceptual evaluations included the assessment of naturalness using a MOS test, and the assessment of system

preference with a pair-wise comparison test. 19 subjects (8 females) participated in the tests. All subjects speak the Hanoi variety of Vietnamese. Participants were 20 to 35 year-old and reported normal hearing. In the test corpus, 40 sentences were chosen so that each sentence covers only one rule, to ease the analysis. These sentences are automatically parsed and then manually corrected at the input of VTed. Three to four examples are designed for each rule.

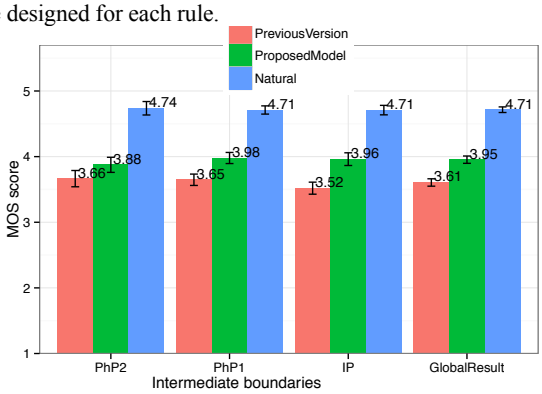

Figure 2: *Results of naturalness using MOS Test.*

The MOS test is carried out with two versions of VTed and a natural speech reference, presented in a random order. Subjects were asked to score *"5-Excellent, 4-Good, 3-Fair, 2-Poor and 1-Bad"* for the naturalness after listening to an utterance. In the pair-wise comparison test, subjects listened these 40 stimuli, composed of two utterances based on the two version of VTed, separated by a "beep" sound. The order of the two voices in each pair and the order of utterances are presented randomly to the subjects. MOS test's results (Figure 2) show an increase of 0.35 on a 5-point MOS scale, for the new prosodic-informed system (3.95/5), compared to the previous TTS system (3.61/5). The pair-comparison test's results (Figure 3) show a preference in about 70% of the cases for the newly proposed model over the previous version.

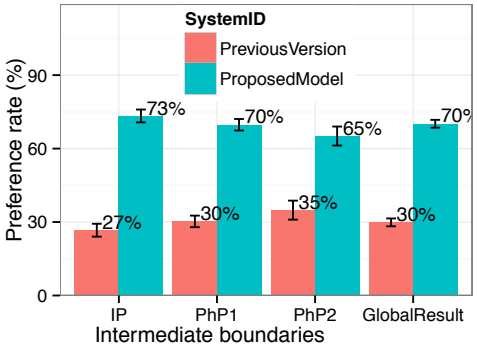

Figure 3: *Results of pair-wise comparison.*

Detailed results for both systems and the proposed syntactic rules and intermediate boundaries are examined. Some rules may highly ameliorate the naturalness (MOS test score >=4.05, rules HC1, HC2, HC6) or the preference (>=81%, rules HC8, HD2, HD3) over the previous version. However, some other ones induce no or little improvement with an increase of only 0.2 in the MOS test or ~49% preference over the previous system (rules HD1, HC3).

Table 7 shows the ANOVA results of the MOS test and the pair-wise comparison test. In the MOS test, the two-factorial ANOVA are the System (3 levels) and the Syntactic Rule (12 levels) or the Break Level (3 levels). In the pair-wise comparison test, a two-factorial ANOVA was run on the results to see if there was a difference in two versions of VTed, according to the two factors Syntactic Rule (12 levels) and Break Level (3 levels).

Table 7: *Anova results of MOS test and pair-wise comparison*

Test	Factor	df	df error	F	p	η^2
MOS score	System	2	2415	806.6	0.000	0.40
	Rule	11	2415	3.0	0.001	0.01
	System:Rule	22	2415	2.0	0.004	0.02
MOS score	System	2	2442	795.0	0.000	0.39
	BreakLevel	4	2442	1.8	0.168	0.00
	System:BreakLevel	4	2442	2.6	0.035	0.00
Preference	Rule	11	805	4.0	0.000	0.05
	BreakLevel	2	814	1.6	0.199	0.00

In the MOS test, the System factor has a significant effect ($p<0.0001$) on the MOS score, and explains an important part of the variance (partial $\eta^2=0.40$). The Syntactic Rule factor ($p<0.0001$) in both tests and its interaction with System factor on MOS score ($p<0.005$) have a significant effect, but show small effect strength. The effect of the Break Level factor and its interaction with System factor on MOS score is not significant.

6 Conclusions and perspectives

Prosodic analysis results confirm that intonation phrase boundaries (i.e. major breaks) appear at the end of a clause having at least 5 syllables or between predicates/head elements, whereas other rules between phrases/words or short clauses can identify 2 levels of phonological phrase boundaries (i.e. minor breaks). This confirmation mainly depends on the logarithm of pause length due to a non-significant effect of boundaries on final lengthening. Break levels (including ones predicted by syntactic rules) and syllable positions relative to prosodic phrases are used to train VTed, an HMM-based TTS system for Vietnamese. In the synthesis phase, break levels (2-4) are explicitly inserted whereas lengthening is applied for the last syllables (linked to syllabic structures and tone types) of prosodic phrases. Perception tests show an average increase of 0.34 on a 5 point MOS scale, for the new prosodic-informed system (3.95/5) compared to the previous one (3.61/5). In the pair-wise comparison test, the new version is preferred with a 70% vs. 30% ratio over the previous version.

The results of ANOVA and post-hoc Tukey tests show that the Break Level factor received a significant effect but explains a small part of variance while the effect of Syntactic Rule factor was not significant. These results raise a need to spend more analysis effort on a larger corpus to understand the origin of noise dispersing the perception results. Some features may be considered during corpus design not only at the phonemic level but also at the syllable level (e.g. syllable structures and tone types of the last syllables in predicted phrases), and at the syntactic level (e.g. more systematic syntactic rules, number of syllables in syntactic elements).

7 Acknowledgements

This work was supported by the Région Ile-de-France through the FUI ADN-TR project (2011-2014), and by the Eiffel Excellence Scholarship. The authors would like to thank Marc Brunelle for interesting discussions on the topic of Prosodic Hierarchy, and last but not least, the subjects who gave time and effort for the experiments.

8 References

[1] F.-C. Chou, C. yu Tseng, and L.-S. Lee, "Automatic generation of prosodic structure for high quality Mandarin speech synthesis.," in *ICSLP*, 1996.

[2] J. Tao, H. Dong, and S. Zhao, "Rule learning based Chinese prosodic phrase prediction," in *2003 International Conference on Natural Language Processing and Knowledge Engineering, 2003. Proceedings*, 2003, pp. 425–432.

[3] D. Doukhan, A. Rilliard, S. Rosset, and C. d' Alessandro, "Modelling pause duration as a function of contextual length.," in *INTERSPEECH*, 2012.

[4] J. Apel, F. Neubarth, H. Pirker, and H. Trost, *Have a break! Modelling pauses in German speech*. 2004.

[5] P. Chistikov and O. Khomitsevich, "Improving Prosodic Break Detection in a Russian TTS System," in *Speech and Computer*, Springer, 2013, pp. 181–188.

[6] O. Jokisch, H. Kruschke, and R. Hoffmann, "Prosodic reading style simulation for text-to-speech synthesis," in *Affective Computing and Intelligent Interaction*, Springer, 2005, pp. 426–432.

[7] A. Parlikar, "Style-Specific Phrasing in Speech Synthesis," Carnegie Mellon University, 2013.

[8] E. O. Selkirk, *On prosodic structure and its relation to syntactic structure*. Indiana University Linguistics Club, 1980.

[9] E. Selkirk, "The Syntax-Phonology Interface," in *The Handbook of Phonological Theory*, J. Goldsmith, J. Riggle, and A. C. L. Yu, Eds. Wiley-Blackwell, 2011, pp. 435–484.

[10] M. Nespor and I. Vogel, "Prosodic Structure Above the Word," in *Prosody: Models and Measurements*, D. A. Cutler and D. D. R. Ladd, Eds. Springer Berlin Heidelberg, 1983, pp. 123–140.

[11] B. Hayes, "The prosodic hierarchy in meter," *Phonetics and phonology*, vol. 1, pp. 201–260, 1989.

[12] N. Dehé, I. Feldhausen, and S. Ishihara, "The prosody–syntax interface: Focus, phrasing, language evolution," *Lingua*, vol. 121, no. 13, pp. 1863–1869, Oct. 2011.

[13] H. A. Viet, D. T. P. Thu, and H. Q. Thang, "Vietnamese Parsing Applying The PCFG Model," in *Proceedings of the Second Asia Pacific International Conference on Information Science and Technology, Vietnam*, 2007.

[14] P.-T. Nguyen, X.-L. Vu, T.-M.-H. Nguyen, V.-H. Nguyen, and H.-P. Le, "Building a Large Syntactically-annotated Corpus of Vietnamese," in *Proceedings of the Third Linguistic Annotation Workshop*, Suntec, Singapore, 2009, pp. 182–185.

[15] A.-C. Le, P.-T. Nguyen, H.-T. Vuong, M.-T. Pham, and T.-B. Ho, "An Experimental Study on Lexicalized Statistical Parsing for Vietnamese," in *Proceedings of the 2009 International Conference on Knowledge and Systems Engineering*, Hanoi, Vietnam, 2009, pp. 162–167.

[16] H. L. T. Bich, N. T. Hung, and H. Ikeda, "Development of Vietnamese Parser based on Phrase Pattern Grammar," in *Proceedings of The Association for Natural Language Processing - 18th Annual Meeting*, Japan, 2012, pp. 947–950.

[17] W. N. Campbell, "Syllable-based segmental duration," *Talking machines: Theories, models, and designs*, pp. 211–224, 1992.

[18] N. Campbell, "Automatic detection of prosodic boundaries in speech," *Speech communication*, vol. 13, no. 3, pp. 343–354, 1993.

[19] T. T. T. Nguyen, C. Alessandro, A. Rilliard, and D. D. Tran, "HMM-based TTS for Hanoi Vietnamese: issues in design and evaluation," *Interspeech 2013*, Lyon, France, Aug-2013.

[20] T. T. T. Nguyen, D. D. Tran, A. Rilliard, C. Alessandro, and T. N. Y. Pham, "Intonation issues in HMM-based speech synthesis for Vietnamese," *The 4th International Workshop on Spoken Language Technologies for Under-resourced Languages (SLTU'14)*, St Petersburg, Russia, May-2014.

[21] 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.

[22] R. D. V. Valin, *An Introduction to Syntax*. Cambridge University Press, 2001.

[23] P. Kroeger, *Analyzing Grammar: An Introduction*. Cambridge University Press, 2005.

[24] M. Nespor and I. Vogel, *Prosodic phonology: with a new foreword*, vol. 28. Walter de Gruyter, 2007.

[25] V.-B. Le and L. Besacier, "Automatic Speech Recognition for Under-resourced Languages: Application to Vietnamese Language," *Trans. Audio, Speech and Lang. Proc.*, vol. 17, no. 8, pp. 1471–1482, Nov. 2009.

[26] 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.

[27] 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.

[28] T. T. T. Nguyen, "VTed: an HMM-based TTS system for Vietnamese.," 2013. [Online]. Available: http://mica.edu.vn/tts3/.