Prosodic Annotation in a Thai Text-to-speech System*

Siripong Potisuk

Department of Electrical and Computer Engineering
The Citadel, The Military College of South Carolina
171 Moultrie Street
Charleston, South Carolina 29409 USA
siripong.potisuk@citadel.edu

Abstract. This paper describes a preliminary work on prosody modeling aspect of a text-to-speech system for Thai. Specifically, the model is designed to predict symbolic markers from text (i.e., prosodic phrase boundaries, accent, and intonation boundaries), and then using these markers to generate pitch, intensity, and durational patterns for the synthesis module of the system. In this paper, a novel method for annotating the prosodic structure of Thai sentences based on dependency representation of syntax is presented. The goal of the annotation process is to predict from text the rhythm of the input sentence when spoken according to its intended meaning. The encoding of the prosodic structure is established by minimizing speech disrhythmy while maintaining the congruency with syntax. That is, each word in the sentence is assigned a prosodic feature called strength dynamic which is based on the dependency representation of syntax. The strength dynamics assigned are then used to obtain rhythmic groupings in terms of a phonological unit called foot. Finally, the foot structure is used to predict the durational pattern of the input sentence. The aforementioned process has been tested on a set of ambiguous sentences, which represents various structural ambiguities involving five types of compounds in Thai.

Keywords: Text-to-speech, Prosody.

1. Introduction

Presently, widespread use of text-to-speech technology is limited by its inability to produce high-quality speech. That is, intelligibility and naturalness of synthetic speech is still not quite at the level acceptable by human listeners. In particular, the naturalness issue can be attributed to the lack of sophisticated prosody-generating scheme.

Prosody is often described as a suprasegmental feature of speech (a term for describing phonological features of those aspects of speech that involve more than single consonants or vowels). Acoustically speaking, prosody can be defined as change in the fundamental frequency (F_0), timing, and amplitude of a speech signal. Speakers control the prosody of an utterance in order to signal linguistic and affective information. Linguistic prosody is used by speakers to signal

* The author would like to thank the Citadel Foundation for its financial support in the form of a presentation grant.

grammatical information at the syllable, word, or sentence level (e.g., stress, intonation). Affective prosody, on the other hand, is used to convey information that indicates speaker's intentions, attitudes, or emotional states. In addition to linguistic and affective information, prosody can also be used to convey non-linguistic information concerning speaker's personal characteristics such as age, gender, idiosyncrasy, speaking style, and physical condition. Such characteristics may or may not be under the speaker's volitional control. It is part of the intelligibility and naturalness of his/her speech. This paper will deal with linguistic prosody only.

The role of linguistic prosody in spoken language is similar to that of punctuation in written language. Punctuation is used to divide a stream of text into smaller segments such as a phrase, clause, or sentence, and thus, helps readers interpret the message according to the intention of the writer. Likewise, prosodic information helps listeners interpret a spoken utterance in the way the speakers intends. The need for punctuation or prosody can be attributed in part to the inherent ambiguity of natural language.

Intuition tells us that intelligibility and naturalness of speech can be attributed to prosody. Some words in an utterance are louder and longer than others. Because function words are acoustically less prominent than the semantically important content words, such as nouns and verbs, we can prosodically distinguish them. Pauses tend to be inserted at certain points in the utterance, and words at the end of the utterance are likely to be lengthened. This suggests the existence of prosodic constituents that are used in the overall prosodic structure or melody of an utterance. Linguists have posited units such as syllables, prosodic words, phonological phrases, and intonational phrases.

The use of prosody by speakers, in attempting to sound intelligibly and naturally, can be best exemplified by considering its use in ambiguous sentences. When two sentences are segmentally identical, a problem of identifying the correct meaning arises for listener, especially when the contextual information is not adequate. In such cases, the listener can make use of another type of information, namely prosody. The question arises, from the speaker's point of view, as to how this information is decoded or associated with different meanings. At an abstract level, a commonly accepted hypothesis is that there is a direct relationship between the syntactic structure of a sentence and its prosodic structure as suggested by Selkirk (1984), and Nespor (1986). This hypothesis implies that an ambiguous sentence will have a different prosodic structure for each syntactic structure, and as such it can be used to determine the correct meaning. At the phonetic level, the speaker tends to manipulate the acoustic correlates of prosody, such as F_0, segmental and pause duration, amplitude, and spectrum of the speech signal in order to signal prosody. The listener, in turn, will try to translate changes in these physical correlates into abstract linguistic concepts in order to arrive at the intended meaning of the utterance.

As in human speech, it is believed that prosodic information can help improve performance of a text-to-speech system. Prosodic information is particularly helpful in generating synthetic speech because of lexical and structural ambiguities of written forms. Prosodic information could be used by computers to generate phonetically similar, but syntactically different utterances.

In the following sections, a novel method for annotating prosody in a text-to-speech system will be described. The process will be abstractly described and demonstrated by using structurally ambiguous sentences involving different types of compounds in Thai. Vongvipanond (1993) concluded that compounds are a major cause of structural ambiguity in Thai and often create problems because of their high frequency of occurrence. Compounding is the most widespread word formation process in Thai. Structural ambiguities often result from compounds because Thai words lack inflectional and derivational affixes to indicate, for example, subject-verb agreement. Nevertheless, compounds can be prosodically distinguished from syntactic phrases by differences in stress patterns. In addition, the process of generating durational patterns for the utterance based on the prosodic annotation process will also be described.

2. Text Processing

Text processing is considered one of the many important aspects of a text-to-speech system involving language modeling. A language model often consists of a grammar written using some formalism which is applied to a sentence by utilizing some sort of parsing algorithm. One popular example is a set of context-free grammar (CFG) production rules, which is based on a phrase-structure representation of syntax by Chomsky (1963), can be used to parse sentences in the language defined by that grammar. A phrase-structure grammar uses a phrase-structure tree (PS-tree) to describe the groupings of words into the so-called *constituents* at different levels of sentence construction. A PS-tree shows which items go together with other items to form tight units of a higher-order, a distributional characteristic of a grouping within a larger grouping. Syntactic class membership is a way of labeling syntactic roles in a PS-tree because a PS-tree does not and cannot specify the types of syntactic links existing between two items in a natural and explicit way. Another approach to syntactic parsing is based on dependency grammar. A dependency grammar describes the syntactic structure of a sentence by using a dependency tree (D-tree) to establish dependencies among words in terms of head and dependents. A D-tree shows a relational characteristic of the syntactic representation in the form of hierarchical links between items, i.e., which items are related to which other items and in which way. In contrast to PS-tree, class membership is not specified in a D-tree. Instead, a D-tree puts a particular emphasis on specifying in detail the type of any syntactic relation between two related items. Such syntactic relations are, for example, predicative, determinative, coordinative relations, etc.

From the above contrastive description of the two approaches to representing the syntax of natural languages, one can draw the following conclusion. The phrase-structure representation is suitable for languages like English, which have a rigid word order and a near absence of syntactically driven morphology. On the other hand, the dependency representation is suitable for languages like Latin or Russian, which feature an incredibly flexible (but far from arbitrary) word order and very rich systems of morphological markings. Word arrangements and inflectional affixes are obviously contingent upon relations between words rather than upon constituents.

In this paper, we argue for the choice of dependency representation of grammar for Thai. We also adopted an alternative formalism, a constraint dependency grammar (CDG) proposed by Potisuk (1996). Thai is the official language of Thailand, a country in the Southeast Asia region. The language is spoken by approximately 65 million people throughout different parts of the country. The written form is used in school and all official forms of communication.

We believe that a CDG parser appears to be an attractive choice for analyzing Thai sentences considering vantage points from both written and spoken language processing aspects of an automatic system. CDG parsers rule out ungrammatical sentences by propagating constraints. Constraints are developed based on a dependency-based representation of syntax. The motivation for our choice of dependency grammar, instead of phrase-structure grammar, stems from the fact that it appears that Thai syntax might be better described by the former representation.

Difficulties in parsing Thai sentences using traditional CFG parsers arise for the following reasons. First, Thai sentences do not contain delimiters or blanks between words. Unlike English, Thai words in a sentence are not flanked by a blank space. Words are concatenated to form a phrase or sentence without explicit word delimiters. This creates a problem for the syntactic analysis of Thai sentences because most parsers operate on words as the smallest syntactic unit in a sentence. To overcome this problem, a word segmentation module must be added to the front end of the parser. This solution, in turn, creates a new problem. Instead of analyzing a single sentence, a parser must now analyze multiple sentence hypotheses comprising a combination of all possible words generated by the word segmentation algorithm. Secondly, Thai words lack inflectional and derivational affixes. Since words in Thai do not inflect to indicate their syntactic function, the position of a word in a sentence alone shows its syntactic function. Hence, syntactic relationships

are primarily determined by word order, and structural ambiguity often arises. Thirdly, inconsistent ordering relations within and across phrasal categories characterize Thai sentences. While a noun, the head of a noun phrase, always precedes its modifying adjectives and determiners, the verb phrase exhibits less consistency. Although a verb, the head of the verb phrase, always precedes its object, its modifying auxiliaries can either precede or follow it. In addition, constituents that optionally occur with the head in both noun and verb phrases, such as determiners and quantifiers, tend to be less consistent in their ordering as well. Lastly, Thai sentences sometimes contain discontinuous sentence constituents in their construction. In grammatical analysis, discontinuity refers to the splitting of a construction by insertion of another grammatical unit. In other words, discontinuity occurs when the elements which make up the constituents are interrupted by elements of another constituent in a sentence.

Given the aforementioned properties of Thai sentences, a CDG parser offers many advantages over traditional CFG parsers in order to overcome the difficulties in parsing Thai. For one thing, CDG is capable of efficiently analyzing free-order languages because order between constituents is not a requirement of the grammatical formalism. Since Thai exhibits significant word order variation, using CFG to describe Thai is cumbersome because numerous rules would be needed to cover all possible configurations of a constituent. Secondly, The CDG approach provides a uniform mechanism of constraint propagation for each knowledge source, i.e., lexical, syntactic, semantic, and pragmatic information, in resolving ambiguities during parsing. The constraints for each knowledge source can be independently developed and applied. A CFG parser, on the other hand, does not provide a good coordinating scheme because it is incapable of selectively invoking different knowledge sources. Concerning the need to analyze multiple sentence hypotheses, our CDG parser allows efficient processing in the form of a constraint network consisting of a directed acyclic word graph augmented with parse-related information. Multiple sentence hypotheses are thus processed simultaneously by pruning the network through the propagation of various constraints. The network also provides a much better representation than a list of sentence hypotheses because it reduces redundancy and compactly represents the set of sentence hypotheses, thereby reducing the storage requirement. Due to the scope of the paper, a description of the basics of CDG parsing of Thai will be omitted. Interested readers are referred to the paper by Potisuk (1996) for a discussion of the basic framework and a parsing example. After all the constraints are propagated across the constraint network and filtering is completed, the network provides a compact representation of all possible parses. Syntactic ambiguity is easy to spot in the network. If multiple parses exist, then additional constraints, such as semantic constraints, can be propagated to further refine the analysis to the intended meaning of the input sentence. The resulting parse trees are then ready to be prosodically annotated. The annotation process is described next.

3. Prosodic Annotation

Prosodic annotation or encoding provides to the prosody-generating module in a text-to-speech system relevant information that adequately captures the essence of the prosodic structure of the input sentence or text. Prosodic encoding usually involves the process of predicting prosodic labels for the input sentence according to the intended meaning. The labeling criteria provide a mechanism for mapping abstract prosodic labels into a sequence of acoustic correlates of prosody. As a result, prosodically-labeled sentences contain information concerning the correspondence between the phonological and phonetic attributes of the prosodic structure of utterances and their intended meanings. Prosodic labels should be chosen to represent abstract linguistic categories of prosody, such as rhythmic groupings (or phrasing) and prominence. Also, they should be chosen such that they are used consistently within and across human labelers, and they make the automatic labeling process tractable and consistent. An example of a prosodic labeling system for English speech is described next.

Price et al. (1991) proposed a labeling system consisting of seven labels, called prosodic break indices. These break indices express the degree of perceived decoupling or separation between every pair of words in an utterance. A boundary within a clitic group (e.g. determiner-noun, two-word verb, etc.) is indicated by a 0 break index; a normal word boundary by a 1; a boundary marking a minor grouping of words by a 2; an intermediate phrase boundary by a 3; an intonational phrase by a 4; a boundary marking a grouping of intonational phrases by a 5; and a sentence boundary by a 6. In terms of prominence, prominent syllables in an utterance are indicated by P1 for a major phrasal prominence; P0 for a lesser prominence; C for contrastive stress; and S for syllables with no prominence. Price demonstrated that these metrics could be used effectively by human labelers to determine how speakers encode prosodic cues for structural ambiguities in structurally ambiguous sentences.

In this paper, we modified the Price's methods in the development of our prosodic encoding scheme for Thai to accommodate the use of dependency grammar formalism. The encoding of the prosodic structure is accomplished by annotating each word in the sentence with a prosodic feature called strength. We describe next how the strength features are derived and compare them with Price's break indices.

The strength feature is chosen based on the dependency representation of syntax. According to the congruency model of syntax and prosody used by Bailly (1983), a relation of dominance between two adjacent lexical items can be established based on their positions in the D-tree. Figure 1 illustrates the four basic configurations of relational marks between adjacent lexical items in a D-tree. ID or independence indicates no direct link between the two items; IT or interdependence indicates the dependence of the two lexical items on the same governor; LD or left dependence indicates the dependence on the following word; RD or right dependence indicates the dependence on the preceding word. It is noted that LD and RD are relational marks between two lexical items at different levels of the D-tree while ID and IT are at the same level.

In addition, we have developed a new set of relational marks called strength dynamics in order to take into account the information about the lexical category of each word in addition to its position in the D-tree. Lexical category information is important because it is related to the stress placement rules in spoken language. Content words are usually stressed; function words are usually unstressed.

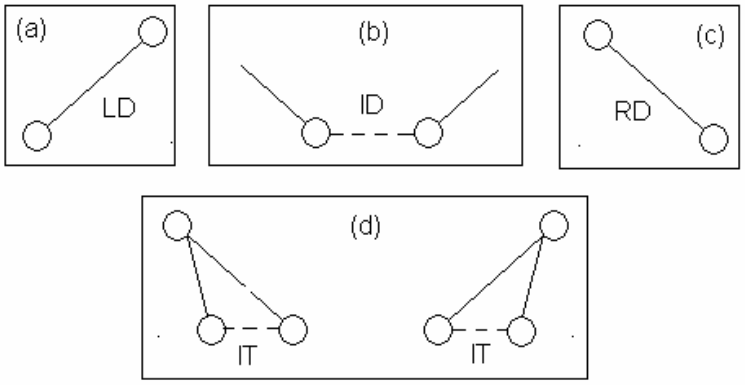

Figure 1: The four basic configurations of the relational marks in the dependency tree: (a) left dependence (LD), (b) independence (ID), (c) right dependence (RD), and (d) interdependence (IT).

There are four levels of strength dynamics: strong dependence (SD), dependence (DE), independence (ID), and strong independence (SI). SD describes a strength dynamic at the word boundary within a clitic group, within a compound, between a content and a function word, or between two function words that are interdependent (i.e., both depend on the same governor). DE describes strength dynamic at minor phrase boundaries, i.e., between a subject noun phrase and a verb phrase, between a verb and an object noun phrase, or between two content words. ID describes strength dynamic at major phrase boundaries (intonational phrases). And, SI describes strength dynamic at the sentence boundary.

Like the break indices used in Price's labeling system, these strength dynamics indicate the degree of connection between the present and the preceding words in an input sentence. They are similar in a sense that both represent the relationship between two adjacent words in a sentence. The stronger the dependency strength is, the smaller the break index. Nonetheless, the strength dynamic has an added benefit in terms of the lexical category information. In addition to the strength feature, a word at the end of a phrase or an utterance will receive the feature '*final*' to indicate that it is affected by the final lengthening effect. Final lengthening is always accompanied by a pause. A word with a '*final*' feature also automatically receives a strength dynamic of ID or SI.

In addition, we utilize a prosodic encoding scheme that integrates both syntactic and rhythmic constraints. That is, the prosodic structure of an utterance is established by minimizing speech disrhythmy while maintaining the congruency with syntax.

Speech is rhythmical not only because of the pattern of sounds and pauses, but also because of the regular recurrence of strongly accented sounds in a series. For example, in a stressed-time language, it is observed by Gee (1983) that speakers tend to produce stressed syllables at a regularly spaced interval of time while they tend to pause according to the syntax of the utterance. The pause distribution seems to be ruled by syntactic constraints. Speech rhythm is also a psychological correlate of speech timing (an objective instrumental measurement of the duration of segments, syllables, etc.) Thus, in a stressed-time language, stress, pause and relative syllable durations interact to form speech rhythm. In addition, the phonology and syntax of the language affect the description of speech rhythm as well.

Thai has a stress-timed rhythm according to Luangthongkum (1977). This means that stressed syllables in Thai are perceived to be isochronous (i.e., they recur approximately at equal intervals of time). A phonological unit called foot is used to describe rhythmic groupings within an utterance. A foot is one of many prosodic constituents and is an elementary unit of the prosodic structure in addition to a syllable. A foot is neither a grammatical nor a lexical unit. The domain of a foot extends from a salient (stressed) syllable up to but not including the next salient syllable. A pause is considered a salient syllable, and the beginning of an utterance is always preceded by a pause. It should be noted that a rhythmic pause has a syntactic function, but a disfluency or hesitation pause does not.

In her analysis of Thai rhythm, Luangthongkum posited five foot structures: | S | = one-syllable foot, | S W | = two-syllable foot, | S W W | = three-syllable foot, | S W W W | = four-syllable foot, | S W W W W | = five-syllable foot, where S and W indicate salient (stressed) and weak (unstressed) syllables, respectively. The four-syllable and five-syllable feet are very rare and are omitted from further discussion. Note that foot boundaries are usually inserted in front of the salient syllables.

Based on the discussion above, the strength dynamics assigned earlier can be used to obtain the information about the foot structure using the following rules. Since we only distinguish between two classes of stress, the salient syllable immediately after a weak syllable receives a strength dynamic of SD. A word before a pause receives a strength dynamic of DE as well as the '*final*' feature. A word after a pause receives a strength dynamic of SI if it is in the utterance-initial position; otherwise, it receives a strength dynamic of ID.

4. Prediction of Durational Patterns

First, we describe the criteria for obtaining duration and pause information from the above strength features (through the derived foot structure). These criteria establish the correspondence between the phonological (strength dynamics) and the phonetic (acoustic correlates) attributes of prosody.

At an abstract level, Luangthongkum assumed that each rhythmic foot is arbitrarily three units long, regardless of the number of syllables comprising the foot. This suggests that as the number of unstressed syllable in the interval increases, a tendency toward equality on inter-stress intervals causes both the stressed and unstressed syllables to become shorter. Thus, the relative syllable duration for each type of rhythmic foot can be abstractly described as follows:

| S | → | 3|
| S W | → | 2 : 1 |
| S W W | → | 1½ : ¾ : ¾ |.

Phonetically, a rhythmic foot is not isochronous. The duration of a foot will differ somewhat depending upon the phonetic structure of the syllables comprising it. Thus, the acoustic realization of a rhythmic foot will be different from the above abstract description. The following is a set of rules proposed by Laungthongkum to predict how syllable durations in each type of foot are realized acoustically. The derived or predicted syllable durations were based on her acoustic analysis of read speech.

| 3 | → | 2 | if the foot is in an utterance-initial position.
| 3 | → | 4 | if the foot is in an utterance-final position and it does not have a CVS structure.
| 2 : 1 | → | 2 : 2| if the salient syllable has a CVS structure; or the weak syllable is the first element of a compound that does not have a CVS structure; or both the salient syllable and the weak syllable are function words.
| 1½ : ¾ : ¾ | → | 1⅔ : 1⅔ : 1⅔ | if the salient syllable has a CVS structure; or it is in an utterance-initial position; or it is a function word and the two weak syllables are two function words or a function word and a linker syllable.

The above approach has been tested on a set of ambiguous sentences, which represents various structural ambiguities involving five types of compounds in Thai: noun-noun, noun-propernoun, noun-verb, noun-verb-noun, and verb-noun. Figure 2 depicts the process of predicting durational patterns from strength dynamics for two hypotheses of an ambiguous sentence of the type noun-verb compound, / **k□□ŋphèt** mâak paj /.

Table 1 lists all types of ambiguous test sentences. There are two test sentences for each type of ambiguity resulting in a total of 10 sentence types for the whole set. These sentences are composed of only monosyllabic words. No polysyllabic words were used because structural ambiguity in Thai does not usually involve polysyllabic words.

5. Conclusion

We have described our preliminary work on prosody modeling to improve intelligibility and naturalness of synthetic speech produced by a Thai text-to-speech system. Such improvement will undoubtedly make this type of speech technology more attractive and acceptable to human listeners. This paper describes the prosody annotation process in which the foot structure (the rhythm of the utterance) is obtained from text. The derived foot structure is then used to predict the durational pattern of the utterance. This prediction of prominence and phrasing patterns from text in general

only operates on single sentences. Whether this technique can be extended to a different prosodic level as in conversational, discourse, or spontaneous communication remains the subject of future investigation. Modeling the discourse effects of prosody is inherently a difficult problem because of a high level of variability in speaker's choices. Furthermore, a design of robust system for describing discourse prosody would not be considered important unless speech synthesis is used in more conversational applications instead of an interaction involving simple questions and declarative sentences.

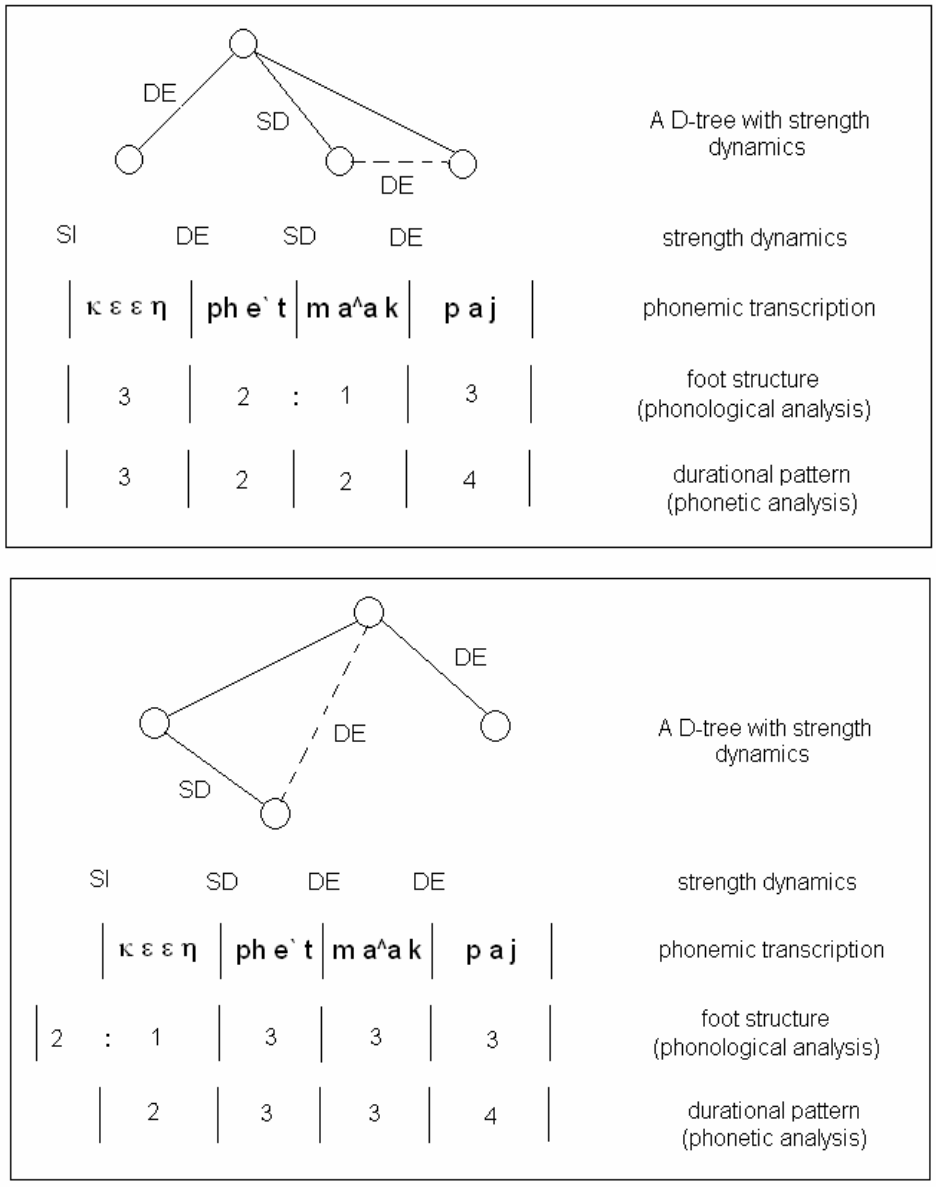

Figure 2: A prediction of durational patterns for two sentence hypotheses of an ambiguous sentence, / **k□□ŋ phèt** mâak paj /. The top panel indicates the first interpretation, 'the curry is too spicy'. The bottom panel indicates the second interpretation, 'there is too much curry.'

Table 1: A list of ambiguous test sentences used for testing the prosodic annotation scheme. The first pair represents a noun-verb compound; the second, a noun-propernoun compound; the third, a noun-noun compound; the fourth, a noun-verb-noun compound; and the fifth, a verb-noun compound.

Thai script	Phonemic transcription	English translation
1. (a) □□□□□□□□□□□□	/ k□□ŋ phèt mâak paj /	'The curry is too spicy.'
(b) □□□□□□□□□□□□	/ k□□ŋphèt mâak paj /	'There is too much curry.'
2. (a) □□□□□□□□□□□□□□	/ phîi nŭu làp jùu /	'Nuu's sister is sleeping.'
(b) □□□□□□□□□□□□□□	/ phîinŭu làp jùu /	'Sister Nuu is sleeping.'
3. (a) □□□□□□□□□□□□□□□□□□□□	/ lûuk lăan maa jiâm b□ □j /	'Our great grandchildren came to visit us quite often.'
(b) □□□□□□□□□□□□□□□□□□□□	/ lûuklăan maa jiâm b□ □j /	'Our children and grandchildren came to visit us quite often.'
4. (a) □□□□□□□□□□□□□□	/ khon khàp rót cháa mâak /	'People drive too slowly.'
(b) □□□□□□□□□□□□□□	/ khonkhàprót cháa mâak /	'The chauffeur was too slow.'
5. (a) □□□□□□□□□□□□□□	/ ph□ˆ□ khàat thun b□ □j /	'Father often runs out of capital.'
(b) □□□□□□□□□□□□□□	/ ph□ˆ□ khàatthun b□ □j /	'Father often suffers business loss.'

References

Bailly, G. 1983. Integration of Rhythmic and Syntactic Constraints in a Model of Generation of French Prosody. *Speech Communication*, 8, 137-146.

Chomsky, N. and M. P. Schutzenberger. 1963. The Algebraic Theory of Context-free Languages. In P. Braffort and D. Hirschberg, eds., *Computer Programming and Formal Systems, Studies in Logic Series,*. 119-161. North-Holland, Amsterdam.

Gee, J. P. and F. Grosjean. 1983. Performance Structures: A Psycholinguistic and Linguistic Appraisal. *Cognitive Psychology*, 15, 411-458.

Luangthongkum, T. 1977. *Rhythm in Standard Thai*. Ph.D. thesis, University of Edinburgh.

Nespor, M. and I. Vogel. 1986. *Prosodic Phonology.* Dordrecht_Holland: Foris.

Potisuk, S. and M. P. Harper. 1996. CDG: An Alternative Formalism for Parsing Written and Spoken Thai. *Proceedings of the Fourth International Symposium on Languages and Linguistics*, pp. 1177-1196.

Price, P., M. Ostendorf, S. Shattuck-Hufnagel and C. Fong. 1991. The Use of Prosody in Syntactic Disambiguation. *Journal of Acoustical Society of America*, 90(6), 2956-2970.

Selkirk, E. O. 1984. *Phonology and Syntax: The Relation between Sound and Structure*. MIT Press.

Vongvipanond, P. E. 1993. Linguistic Problems in Computer Processing of the Thai Language. *Proceedings of the Symposium on Natural Language Processing in Thailand*, pp. 519-545.