USING STRESS TO DISAMBIGUATE SPOKEN THAI SENTENCES CONTAINING SYNTACTIC AMBIGUITY[1]

Siripong Potisuk[2], Mary P. Harper[2], and Jack Gandour[3]

[2] School of Electrical and Computer Engineering
[3] Department of Audiology and Speech Sciences
Purdue University, West Lafayette IN 47907
Email: pong@sanders.ecn.purdue.edu, harper@ecn.purdue.edu, gandour@sage.cc.purdue.edu

ABSTRACT

We have developed a Bayesian classifier to determine whether syllables in connected Thai speech are weakly or strongly stressed by using five acoustic parameters: syllable rhyme duration, mean F_0, F_0 standard deviation, mean energy, and the standard deviation of the energy. With speaker-dependent data normalization, we achieved a classification accuracy of 99%. The classification accuracy drops to 96% when we used speaker-independent normalization. We have also developed prosodic constraints that can use this stress information to syntactically disambiguate a class of ambiguous sentences that arise from the use of compounding in spoken Thai.

1. INTRODUCTION

A major goal of our research is to utilize the prosodic features of spoken Thai sentences to help determine the intended meaning in an automatic speech understanding system. There are two aspects of incorporating prosody into this system: the automatic detection and classification of prosodic cues from the speech signal, and the incorporation of a prosodic disambiguation process into the language model.

Thai belongs to a class of languages called tone languages in which variations in F_0 at the syllable level signal differences in lexical meaning. In the past, we have examined the problem of tone classification. This paper focuses on the problem of stress detection and the use of this feature to disambiguate structurally ambiguous Thai sentences.

Thai is considered an isolating or analytic language in which sentences are constructed from sequences of free morphemes, each word consisting of a single morpheme. Semantic and grammatical concepts (e.g., tense) are expressed through the use of free morphemes rather than a change of form or affix. Compounds in Thai are very important not only because of their high frequency of occurrence, but because they provide us with a window to see how prosody can be used by listeners to resolve ambiguities in Thai. Compounds can be distinguished from other syntactic phrases by differences in stress patterns. For monosyllabic words in Thai, all content words are strongly stressed while all grammatical words or clitics are weakly stressed (unless reinforced by emphasis). Hence, a bisyllabic noun compound has a weak-strong stress pattern compared with a subject-verb construct which has a strong-strong stress pattern.

We have developed a Bayesian classifier, which is described in section 2, to determine whether syllables in connected Thai speech are weakly or strongly stressed. Then in section 3, we describe how this stress information can be used to disambiguate syntactically ambiguous sentences that result from the use of compounding in Thai.

2. STRESS CLASSIFICATION ALGORITHM

The problem of stress classification can be described in terms of a statistical pattern recognition system. Traditionally, the design of a statistical pattern recognition system involves two major steps: feature extraction and pattern matching. The feature extraction step produces a sequence of feature vectors representing a set of measurements associated with a frame of the physical signal of interest. The statistical pattern matching step involves determining an appropriate statistical model and selecting the maximum likelihood classification corresponding to the sequence of feature vectors such that classification error is minimized. Following [19], we use a Bayesian classifier with only two classes, weak and strong stress. We assume that the acoustic features can be modeled by a normally-distributed population for each class.

2.1. Classification Features

Five acoustic parameters were chosen as classification features: syllable rhyme duration, mean F_0, F_0 standard deviation (SD), mean energy, and energy SD. It is hypothesized that these five acoustic features adequately capture the acoustically discriminatory information for determining syllable stress. The mean and standard deviation of normalized F_0 trajectories were chosen as parameters to characterize changes in F_0 height and shape, respectively. Likewise, the mean energy and energy SD were chosen as parameters to characterize changes in energy level and the shape of the energy contour, respectively. Rhyme duration was chosen because it has been shown to be a better correlate of stress

[1] This work has been supported in part by NIH under grant number DC00515-07. The first author can now be contacted at: 85/273 Phaholyotin 32, Ladprao, Bangkok 10230, Thailand.

than the nucleus or the entire syllable [17, 18].

2.2. Training and Testing Data

The speech data from an acoustic measurement experiment concerning the acoustic correlates of stress in Thai [13] were used as training and testing data for the stress classification algorithm because that corpus contains an ensemble of different types of syllable structures and tonal combinations. In addition, sentence pairs manifested only one type of structural ambiguity (noun-verb sequence vs. compound noun).

Five native speakers of Thai (four graduate and one undergraduate; three male and two female) were each asked to read 25 pairs of ambiguous sentences at a conversational speaking rate. The two sentences in a pair contained six segmentally identical syllables including a two-syllable sequence that occurred at the beginning of the sentences. Vowel length was held constant within sentence pairs. Sentence pairs manifested one type of structural ambiguity. The first member contained a two-syllable noun-verb sequence exhibiting a strong-strong stress pattern; the second member, a two-syllable noun compound exhibiting a weak-strong stress pattern. The tones of each two-syllable sequence were varied to represent all possible two-tone combinations of five Thai tones. Because of the intended ambiguity, each utterance was preceded by a few sentences of disambiguating context. A total of 1250 utterances were recorded (25 sentence pairs X 2 members X 5 repetitions X 5 speakers). Both members of a sentence pair did not occur in the same recording session; the first member of each pair was assigned to the first session, the second member to the second session.

The tape-recorded stimuli were low-pass filtered at 10 KHz and digitized at a sampling rate of 20 KHz by means of a 16-bit A/D converter with a 5-V dynamic range using the KAY CSL (Computerized Speech Lab) Model 4300 installed on a Gateway 2000 P5-90 microcomputer. Duration, F_0, and rms energy of the rhyme portion of the target syllable were selected for measurement. Syllable rhyme onset and offset was determined by hand by positioning cursors on the simultaneous CSL display of an audio waveform and a wide-band spectrogram (8 KHz frequency range, 300 Hz bandwidth). Spectrograms were demarcated in time following conventional rules for segmentation of the speech signal. F_0 was computed directly from the waveform using a CSL algorithm that employs a time domain approach to pitch analysis (modified autocorrelation with center clipping) with nonoverlapping variable frame length. The energy calculation in decibels (dB) was performed in a nonoverlapping frame-by-frame, pitch asynchronous manner using a CSL algorithm in which frame length was kept constant at 20 ms for all speakers.

Duration, F_0, and rms energy were subjected to normalization procedures in order to eliminate confounding influences of stress and tonal categories on the evaluation of stress. Data were normalized first before extracting the five acoustic features.

F_0 contours were equalized for duration on a percentage scale to neutralize durational differences that are correlated with the five lexical tones [1, 7, 9, 22]. Differences in excursion size of F_0 movements related to differences in voice range between speakers were normalized by converting raw F_0 values to an equivalent-rectangular-bandwidth-rate (ERB) scale, a psychoacoustic scale that gives equal prominence to excursions in different pitch registers [6]. Differences in absolute height of F_0 movements between speakers were normalized by transforming the ERB values to a z-score scale [16]. By such normalization procedures, the height and shape of F_0 contours can be evaluated across stress categories free of intrinsic differences due to tonal category or a speaker's voice range.

To compensate for differences in speaking rate within and between speakers, rhyme duration was divided by the total utterance duration so that rhyme duration would be expressed as a proportion of the total utterance duration. To neutralize durational differences due to differences in segmental composition of the target syllable, each rhyme duration proportion was subsequently normalized across stress categories to a z-score scale.

The raw energy contours in dB were normalized to a z-score scale within speaker and tone across stress categories to neutralize intrinsic differences in energy due to tonal category [20]. Following Wightman and Ostendorf [21], the mean energy was analyzed in the target syllable itself without normalization to the mean energy in the following syllable or the mean energy in the utterance.

2.3. Results

With the aid of the SAS^{TM} statistical software package, a series of discriminant function analyses was conducted to determine how useful the combined acoustic features were in classifying the stress on syllables. Each analysis was performed with one of the three different combinations of acoustic features: all five features, three features (duration, F_0 SD, and energy SD), and only one feature (rhyme duration). Cross-validation and split-sample techniques were used to minimize sampling error. For the split-sample technique, the odd-numbered speakers (n=3) provided the training data while even-numbered speakers (n=2), the testing data. The split-sample technique represents a speaker-independent testing procedure. The proportion of cases correctly classified was used as a measure of the accuracy of the procedure, and indirectly, the degree of separation between weakly and strongly stressed syllables in Thai. Bivariate correlation coefficients were used to measure the relative contribution of each acoustic parameter on the discriminant function. The discriminant function analyses were performed on the data normalized in both a speaker-dependent and speaker-independent fashion.

The classification results from a series of discriminant function analyses of stress categories performed on the data normalized in a speaker-dependent fashion indicate that the combined acoustic parameters were very useful in classifying syllable stress. The canonical R^2 equals 0.8688, 0.8687, and 0.8627 for a combination of five, three, and one acoustic parameters, respectively. The classification matrices resulting from the analyses revealed that the percentage of correctly classified individual syllables ranges from 98.83% to 100% across stress categories. Nearly identical percent-

ages of correct classification were obtained with the cross-validation (99.2%) and split-sample (99.8%) analyses. For a combination of all five acoustic parameters, Pearson correlation coefficients between each acoustic parameter and the discriminant function showed that the discriminant function was closely related to rhyme duration ($r = 0.99$, $p < .0001$) and F_0 standard deviation ($r = 0.48$, $p < .0001$). The other three acoustic parameters failed to show a close relationship to the discriminant function, suggesting that the stress distinction can largely be attributed to differences in the rhyme duration. Indeed, we found that there is virtually no drop in performance when rhyme duration alone is used (99.2%). The classification results performed on the data normalized in a speaker-independent fashion were similar; however, there was a slight drop in performance due to the normalization procedure (the percentage of correctly classified individual syllables ranges from 95.63% to 100% across stress categories).

This series of discriminant function analyses indicates that the combined acoustic parameters were very useful in classifying syllable stress in Thai. In fact, our stress classification results confirm the results of [13], that duration is the dominant cue in signaling stress in Thai. Indeed, a relatively simple Bayesian classifier utilizing only duration information has been shown to be effective in classifying stress in Thai. However, the above implementation of an automatic stress classifier was subject to a number of limitations, such as the lack of automatic syllable segmentation and a more phonetically, prosodically, and syntactically balanced speech database. Performance degradation is to be expected when such limitations are overcome. It is also noted that whether or not these results can be generalized across various prosodic structures remains the subject of future research.

3. USING PROSODIC INFORMATION

In order for prosodic information to be used in high-level processing, a mechanism for passing up such information from the low-level acoustic module, called prosodic encoding or annotation, must be devised. Prosodic encoding of an utterance usually involves the process of labeling prosodic patterns in the speech signal. The labeling criteria provide a mechanism for mapping sequences of acoustic correlates of prosody into abstract prosodic labels. Prosodic labels should be chosen to represent the abstract linguistic categories of prosody, such as rhythmic groupings (or phrasing) and prominence, and such that they are used consistently by human labelers. Price et al. [15] proposed a labeling system consisting of seven labels, called prosodic break indices, which express the degree of perceived decoupling or separation between every pair of words in an utterance. Based on this labeling system, Wightman and Ostendorf [21] developed an algorithm for automatically generating the prosodic labels.

There are several ways to use the information provided by prosodic encoding to resolve ambiguity. One approach is to use the prosodic information directly in the rules of the parser [3, 11]. This complicates grammar rules and makes them difficult to design and modify. Another approach involves the post-parsing operation of computing the prosodic consistency score for each parse. This post-parsing method can be easily combined with any parser because it does not directly affect grammar rules of the parser [8, 12]. The disadvantage of this approach is that the prosodic knowledge source is isolated from the other high-level knowledge sources.

Rather than use either of these approaches, we add prosodic constraints as an additional knowledge source to our constraint dependency grammar (CDG) parser [4]. CDG parsers rule out ungrammatical sentences by propagating constraints. Constraints are developed based on a dependency-based representation of syntax. It is a simple matter to develop semantic constraints for a specific corpus of sentences [5], and prosodic constraints are as easy to develop. In [14], we have discussed the benefits of using constraint dependency grammars for parsing Thai sentences.

We use a method similar to Price's break indices for our prosodic encoding scheme for Thai with a slight modification to take advantage of our CDG parsing framework. The encoding of the prosodic structure is accomplished by annotating each word candidate in a word graph, the central data structure of our constraint-based parser, with a prosodic feature called strength. The strength feature is chosen based on the dependency representation of syntax. A dependency grammar expresses the syntactic relations that lexical items can have with each other using governor-dependent relations in a D-tree. According to the congruency model of syntax and prosody [2], the relation of dominance between two adjacent lexical items can be established based on their positions in the D-tree.

For spoken language parsing, we have developed a set of relational marks called strength dynamics in order to take into account the information about the lexical category of each word candidate along with its position in the D-tree. Strong dependence (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). Dependence (DE) describes a 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. Independence (ID) describes a strength dynamic at major phrase boundaries (intonational phrases). Strong independence (SI) describes a strength dynamic at the sentence boundary.

The labeling criteria establish the correspondence between the phonological (strength dynamics) and the phonetic (acoustic correlates) attributes of prosody. 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. A phonological unit called a foot is used to describe rhythmic groupings within an utterance. The domain of a foot extends from a salient (strongly 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 [10] posited five foot structures.

By using stress, pause, and derived syllable duration information, the input utterance can be divided into feet. Then, the strength dynamics can be assigned as follows. Since we only distinguish between two classes of stress, the salient syllable immediately after a weak syllable receives a strength of SD; otherwise, it receives a strength of DE. The weak syllable receives a strength of SD. A word before a pause receives a strength of DE as well as the final feature. A word after a pause receives a strength of SI if it is in the utterance-initial position; otherwise, it receives a strength of ID.

Syntactic constraints were initially developed for a corpus of ambiguous sentences representing various structural ambiguities involving five types of compounds in Thai: noun-noun, noun-propernoun, noun-verb, noun-verb-noun, and verb-noun. There were 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 because structural ambiguity in Thai does not usually involve polysyllabic words. A set of seven prosodic constraints was developed to distinguish the five types of compounds from other syntactic structures. These prosodic constraints check for the agreement between adjacent pairs of words in each of the competing sentence hypotheses of an ambiguous utterance based on the annotated strength dynamics.

We assume that the word graphs processed by the CDG parser are perfectly annotated with prosodic information (i.e., strength dynamics). Once the word graph for a sentence is annotated with strength dynamics, a combination of syntactic and prosodic constraints are applied to the graph to rule out as many prosodically implausible sentence hypotheses as possible. The parsing process begins by propagating syntactic constraints to eliminate syntactically ill-formed parses. Then, prosodic constraints are propagated to check the agreement between every pair of word candidates in each of the remaining competing parses based on the annotated strength dynamics. A parse is rejected if its syntactic structure is incompatible with the prosodic structure encoded via the annotated strength dynamics. For each ambiguous sentence annotated with strength dynamic information associated with one of the interpretations, the prosodic constraints were able to select the correct parse.

The implementations of the prosodic processing algorithms described in this paper are still at an early stage of development. These algorithms must be further automated to more accurately measure the quality of our approach. Nevertheless, we believe that we have demonstrated that stress in Thai can be measured fairly reliably and that strength dynamic information is useful for disambiguating an important class of Thai

4. REFERENCES

1. A. S. Abramson. The vowels and tones of standard Thai: Acoustical measurements and experiments. *International Journal of American Linguistics*, 28-2, 1962.
2. G. Bailly. Integration of rhythmic and syntactic constraints in a model of generation of French prosody. *Speech Communication*, 8:137–146, 1989.
3. J. Bear and P. Price. Prosody, syntax, and parsing. In *Proceedings of the 28th Annual Meeting of the Association of Computational Linguistics*, pages 17–22, 1990.
4. M. P. Harper and R. A. Helzerman. Extensions to constraint dependency grammar. *Computer Speech and Language*, 9:187–234, 1995.
5. M. P. Harper, L. H. Jamieson, C. B. Zoltowski, and R. A. Helzerman. Semantics and constraint parsing of word graphs. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages II–63–II–66, April 1992.
6. D. Hermes and J. Van Gestel. The frequency scale of speech intonation. *Journal of the Acoustical Society of America*, 90:97–102, 1991.
7. J. Howie. On the domain of tone in Mandarin. *Phonetica*, 30:129–148, 1974.
8. A. Hunt. A generalised model for utilising prosodic information in continuous speech recognition. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages II–169–II–172, May 1994.
9. Q-M. Kong. Influence of tones upon vowel duration in Cantonese. *Language and Speech*, 30:387–399, 1987.
10. T. Luangthongkum. *Rhythm in standard Thai*. PhD thesis, University of Edinburgh, 1977.
11. M. Ostendorf, P. Price, J. Bear, and C. W. Wightman. The use of relative duration in syntactic disambiguation. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing*, pages I–13–I–16, 1990.
12. M. Ostendorf, C. W. Wightman, and N. M. Veilleux. Parse scoring with prosodic information: An analysis-synthesis approach. *Computer Speech and Language*, 7:193–210, 1993.
13. S. Potisuk, J. Gandour, and M. P. Harper. Acoustic correlates of stress in Thai. *Phonetica*, to appear.
14. S. Potisuk and M. P. Harper. CDG: An alternative formalism for parsing written and spoken Thai. In *Proceedings of the Fourth International Symposium on Language and Linguistics: Pan-Asiatic Linguistics*, volume 4, pages 1177–1196, 1996.
15. P. Price, M. Ostendorf, S. Shattuck-Hufnagel, and C. Fong. The use of prosody in syntactic disambiguation. *Journal of Acoustical Society of America*, 90-6:2956–2970, 1991.
16. P. Rose. Considerations in the normalisation of the fundamental frequency of linguistic tone. *Speech Communication*, 6:343–351, 1987.
17. A. M. C. Sluijter and V. J. van Heuven. Effects of focus distribution, pitch accent and lexical stress on the temporal organization of syllables in Dutch. *Phonetica*, 52:71–89, 1995.
18. H. G. Van der Hulst. *Syllable Structure and Stress in Dutch*. Foris, Dordrecht, 1984.
19. A. Waibel. *Prosody and Speech Recognition*. Morgan Kaufmann Publishers, Los Altos, CA, 1988.
20. D. Whalen and Y. Xu. Information for Mandarin tones in the amplitude contours and brief segments. *Phonetica*, 49:25–47, 1992.
21. C. W. Wightman and M. Ostendorf. Automatic labeling of prosodic patterns. *IEEE Transactions on Speech and Audio Processing*, 2:469–481, 1994.
22. E. Zee. Duration and intensity as correlates of F_0. *Journal of Phonetics*, 6:213–220, 1978.