Learning Intonation Rules for Concept to Speech Generation

Shimei Pan and **Kathleen McKeown**
Dept. of Computer Science
Columbia University
New York, NY 10027, USA
{pan, kathy}@cs.columbia.edu

Abstract

In this paper, we report on an effort to provide a general-purpose spoken language generation tool for Concept-to-Speech (CTS) applications by extending a widely used text generation package, FUF/SURGE, with an intonation generation component. As a first step, we applied machine learning and statistical models to learn intonation rules based on the semantic and syntactic information typically represented in FUF/SURGE at the sentence level. The results of this study are a set of intonation rules learned automatically which can be directly implemented in our intonation generation component. Through 5-fold cross-validation, we show that the learned rules achieve around 90% accuracy for break index, boundary tone and phrase accent and 80% accuracy for pitch accent. Our study is unique in its use of features produced by language generation to control intonation. The methodology adopted here can be employed directly when more discourse/pragmatic information is to be considered in the future.

1 Motivation

Speech is rapidly becoming a viable medium for interaction with real-world applications. Spoken language interfaces to on-line information, such as plane or train schedules, through display-less systems, such as telephone interfaces, are well under development. Speech interfaces are also widely used in applications where eyes-free and hands-free communication is critical, such as car navigation. Natural language generation (NLG) can enhance the ability of such systems to communicate naturally and effectively by allowing the system to tailor, reorganize, or summarize lengthy database responses. For example, in our work on a multimedia generation system where speech and graphics generation techniques are used to automatically summarize patient's pre-, during, and post-, operation status to different caregivers (Dalal et al., 1996), records relevant to patient status can easily number in the thousands. Through content planning, sentence planning and lexical selection, the NLG component is able to provide a concise, yet informative, briefing automatically through spoken and written language coordinated with graphics (McKeown et al., 1997) .

Integrating language generation with speech synthesis within a Concept-to-Speech (CTS) system not only brings the individual benefits of each; as an integrated system, CTS can take advantage of the availability of rich structural information constructed by the underlying NLG component to improve the quality of synthesized speech. Together, they have the potential of generating better speech than Text-to-Speech (TTS) systems. In this paper, we present a series of experiments that use machine learning to identify correlation between intonation and features produced by a robust language generation tool, the FUF/SURGE system (Elhadad, 1993; Robin, 1994). The ultimate goal of this study is to provide a spoken language generation tool based on FUF/SURGE, extended with an intonation generation component to facilitate the development of new CTS applications.

2 Related Theories

Two elements form the theoretical background of this work: the grammar used in FUF/SURGE and Pierrehumbert's intonation theory (Pierrehumbert, 1980). Our study aims at identifying the relations between the semantic/syntactic information produced by FUF/SURGE and four intonational features of Pierrehumbert: pitch accent, phrase accent, boundary tone and intermediate/intonational phrase boundaries.

The FUF/SURGE grammar is primarily based on systemic grammar (Halliday, 1985). In systemic grammar, the process (ultimately realized as the verb) is the core of a clause's semantic structure. Obligatory semantic roles, called participants, are associated with each process. Usually, participants convey who/what is involved in the process. The process also has non-obligatory peripheral semantic roles called circumstances. Circumstances answer questions such as when/where/how/why. In FUF/SURGE, this semantic description is unified with a syntactic grammar to generate a syntactic description. All semantic, syntactic and lexical information, which are produced during the generation process, are kept in a final Functional Description (FD), before linearizing the syntactic structure into a linear string. The features used in our intonation model are mainly extracted from this final FD.

The intonation theory proposed in (Pierrehumbert, 1980) is used to describe the intonation structure. Based on her intonation grammar, the F0 pitch contour is described by a set of intonational features. The tune of a sentence is formed by one or more intonational phrases. Each intonational phrase consists of one or more intermediate phrases followed by a boundary tone. A well-formed intermediate phrase has one or more pitch accents followed by a phrase accent. Based on this theory, there are four features which are critical in deciding the F0 contour: the placement of intonational or intermediate phrase boundaries (break index 4 and 3 in ToBI annotation convention (Beckman and Hirschberg, 1994)), the tonal type at these boundaries (the phrase accent and the boundary tone), and the F0 local maximum or minimum (the pitch accent).

3 Related Work

Previous work on intonation modeling primarily focused on TTS applications. For example, in (Bachenko and Fitzpatrick, 1990), a set of hand-crafted rules are used to determine discourse neutral prosodic phrasing, achieving an accuracy of approximately 85%. Recently, researchers improved on manual development of rules by acquiring prosodic phrasing rules with machine learning tools. In (Wang and Hirschberg, 1992), Classification And Regression Tree (CART) (Brieman et al., 1984) was used to produce a decision tree to predict the location of prosodic phrase boundaries, yielding a high accuracy, around 90%. Similar methods were also employed in predicting pitch accent for TTS in (Hirschberg, 1993). Hirschberg exploited various features derived from text analysis, such as part of speech tags, information status (i.e. given/new, contrast), and cue phrases; both hand-crafted and automatically learned rules achieved 80-98% success depending on the type of speech corpus. Until recently, there has been only limited effort on modeling intonation for CTS (Davis and Hirschberg, 1988; Young and Fallside, 1979; Prevost, 1995). Many CTS systems were simplified as text generation followed by TTS. Others that do integrate generation make use of the structural information provided by the NLG component (Prevost, 1995). However, most previous CTS systems are not based on large scale general NLG systems.

4 Modeling Intonation

While previous research provides some correlation between linguistic features and intonation, more knowledge is needed. The NLG component provides very rich syntactic and semantic information which has not been explored before for intonation modeling. This includes, for example, the semantic role played by each semantic constituent. In developing a CTS, it is worth taking advantage of these features.

Previous TTS research results cannot be implemented directly in our intonation generation component. Many features studied in TTS are not provided by FUF/SURGE. For example, the part-of-speech (POS) tags in FUF/SURGE are different from those used in TTS. Furthermore, it make little sense to apply part of speech tagging to generated text instead of using the accurate POS provided in a NLG system. Finally, NLG provides information that is difficult to accurately obtain from full text (e.g., complete syntactic parses).

These motivating factors led us to carry out a study consisting of a series of three experiments designed to answer the following questions:

- How do the different features produced by FUF/SURGE contribute to determining intonation?
- What is the minimal number of features needed to achieve the best accuracy for each of the four intonation features?
- Does intra-sentential context improve accuracy?

```
((cat clause)
 (process ((type ascriptive)
           (mode equative)))
 (participant
      ((identified ((lex "John")
                    (cat proper)))
       (identifier ((lex "teacher")
                    (cat common))))))
```

Figure 1: Semantic description

4.1 Tools and Data

In order to model intonational features automatically, features from FUF/SURGE and a speech corpus are provided as input to a machine learning tool called RIPPER (Cohen, 1995), which produces a set of classification rules based on the training examples. The performance of RIPPER is comparable to benchmark decision tree induction systems such as CART and C4.5. We also employ a statistical method based on a generalized linear model (Chambers and Hastie, 1992) provided in the S package to select salient predictors for input to RIPPER.

Figure 1 shows the input Functional Description(FD) for the sentence " John is the teacher". After this FD is unified with the syntactic grammar, SURGE, the resulting FD includes hundreds of semantic, syntactic and lexical features. We extract 13 features shown in Table 1 which are more closely related to intonation as indicated by previous research. We have chosen features which are applicable to most words to avoid unspecified values in the training data. For example, "tense" is not extracted simply because it can be only applied to verbs. Table 1 includes descriptions for each of the features used. These are divided into semantic, syntactic, and semi-syntactic/semantic features which describe the syntactic properties of semantic constituents. Finally, word position (NO.) and the actual word (LEX) are extracted directly from the linearized string.

About 400 isolated sentences with wide coverage of various linguistic phenomena were created as test cases for FUF/SURGE when it was developed. We asked two male native speakers to read 258 sentences, each sentence may be repeated several times. The speech was recorded on a DAT in an office. The most fluent version of each sentence was kept. The resulting speech was transcribed by one author based on ToBI with break index, pitch accent, phrase accent and boundary tone labeled, using the XWAVE speech analysis tool. The 13 features described in Table 1 as well as one intonation feature are used as predictors for the response intonation feature. The final corpus contains 258 sentences for each speaker, including 119 noun phrases, 37 of which have embeded sentences, and 139 sentences. The average sentence/phrase length is 5.43 words. The baseline performance achieved by always guessing the majority class is 67.09% for break index, 54.10% for pitch accent, 66.23% for phrase accent and 79.37% for boundary tone based on the speech corpus from one speaker. The relatively high baseline for boundary tone is because for most of the cases, there is only one L% boundary tone at the end of each sentence in our training data. Speaker effect on intonation is briefly studied in experiment 2. All other experiments used data from one speaker with the above baselines.

4.2 Experiments

4.2.1 Interesting Combinations

Our first set of experiments was designed as an initial test of how the features from FUF/SURGE contribute to intonation. We focused on how the newly available semantic features affect intonation. We were also interested in finding out whether the 13 selected features are redundant in making intonation decisions.

We started from a simple model which includes only 3 factors, the type of semantic constituent boundary before (BB) and after (BA) the word, and part of speech (POS). The semantic constituent boundary can take on 6 different values; for example, it can be a clause boundary, a boundary associated with a primary semantic role (e.g., a participant), with a secondary semantic role (e.g., a type of modifier), among others. Our purpose in this experiment was to test how well the model can do with a limited number of parameters. Applying RIPPER to the simple model yielded rules that significantly improved performance over the baseline models. For example, the accuracy of the rules learned for break index increases to 87.37% from 67.09%; the average improvement on all 4 intonational features is 19.33%.

Next, we ran two additional tests, one with additional syntactic features and another with additional semantic features. The results show that the two new models behave similarly on all intonational features; they both achieve some

Category	Label	Description	Examples
Semantic	BB	The semantic constituent boundary before the word.	participant boundaries or circumstance boundaries etc.
	BA	The semantic constituent boundary after the word.	participant boundaries or circumstance boundaries etc.
	SEMFUN	The semantic feature of the word.	The semantic feature of "did" in "I did know him." is "insistence".
	SP	The semantic role played by the immediate parental semantic constituent of the word.	The SP of "teacher" in "John is the teacher" is "identifier".
	GSP	The generic semantic role played by the immediate parental semantic constituent of the word.	The GSP of "teacher" in "John is the teacher" is "participant"
Syntactic	POS	The part of speech of the word	common noun, proper noun etc.
	GPOS	The generic part of speech of the word	noun is the corresponding GPOS of both common noun and proper noun.
	SYNFUN	The syntactic function of the word	The SYNFUN of "teacher" in "the teacher" is "head".
Semi-semantic& syntactic	SPPOS	The part of speech of the immediate parental semantic constituent of the word.	The SPPOS of "teacher" is "common noun".
	SPGPOS	The generic part of speech of the immediate parental semantic constituent of the word.	The SPGPOS of "teacher" in "the teacher" is "noun phrase".
	SPSYNFUN	The syntactic function of the immediate parental semantic constituent of the word.	The SPSYNFUN of "teacher" in "John is the teacher" is "subject complement.
Misc.	NO.	The position of the word in a sentence	1, 2, 3, 4 etc.
	LEX	The lexical form of the word	"John", "is", "the", "teacher"etc.

Table 1: Features extracted from FUF and SURGE

improvements over the simple model, and the new semantic model (containing the features SEMFUN, SP and GSP in addition to BB, BA and POS) also achieves some improvements over the syntactic model (containing GPOS, SYNFUN, SPPOS, SPGPOS and SPSYNFUN in addition to BB, BA and POS), but none of these improvements are statistically significant using binomial test.

Finally, we ran an experiment using all 13 features, plus one intonational feature. The performance achieved by using all predictors was a little worse than the semantic model but a little better than the simple model. Again none of these changes are statistically significant.

This experiment suggests that there is some redundancy among features. All the more complicated models failed to achieve significant improvements over the simple model which only has three features. Thus, overall, we can conclude from this first set of experiments that FUF/SURGE features do improve performance over the baseline, but they do not indicate conclusively which features are best for each of the 4 intonation models.

4.2.2 Salient Predictors

Although RIPPER has the ability to select predictors for its rules which increase accuracy, it's not clear whether all the features in the RIPPER rules are necessary. Our first experiment seems to suggest that irrelevant features could damage the performance of RIPPER because the model with all features generally performs worse than the semantic model. Therefore, the purpose of the second experiment is to find the salient predictors and eliminate redundant and irrelevant ones. The result of this study also helps us gain a better understanding of the relations between FUF/SURGE features and intonation.

Since the response variables, such as break index and pitch accent, are categorical values, a generalized linear model is appropriate. We mapped all intonation features into binary values as required in this framework (e.g., pitch accent is mapped to either "accent" or "deaccent"). The resulting data are analyzed by the generalized linear model in a step-wise fashion. At each step, a predictor is selected and dropped based on how well the new model can fit the data. For example, in the break index model, after GSP is dropped, the new model achieves the same performance as the initial model. This suggests that GSP is redundant for break index.

Since the mapping process removes distinctions within the original categories, it is possible that the simplified model will not perform as well as the original model. To confirm that the simplified model still performs reasonably well, the new simplified models are tested by

Model	Selected Features	Dropped features	Model Accuracy		Rule No.		Conditions	
			New	Initial	New	Initial	New	Initial
Break Index	BB BA GPOS SPGPOS SPSYNFUN	NO LEX POS SPPOS SP GSP SEMFUN SYNFUN ACCENT	87.94%	88.29%	7	9	18	16
Pitch Accent	NO BB BA POS GPOS SYNFUN SEMFUN GSP SPPOS SPGPOS SPSYNFUN INDEX	LEX SP	73.87%	73.95%	11	11	20	21
Phrase Accent	NO BB BA POS GPOS SYNFUN SPPOS SPGPOS SPSYNFUN ACCENT	LEX SP GSP SEMFUN	86.72%	88.08%	5	9	15	25
Boundary Tone	NO BB BA GSP	LEX POS GPOS SYNFUN SEMFUN SP SPPOS SPGPOS SPSYNFUN ACCENT	97.36%	96.79%	2	5	4	8

Table 2: The New model v.s. the original model

letting RIPPER learn new rules based only on the selected predictors.

Table 2 shows the performance of the new models versus the original models. As shown in the "selected features" and "dropped features" column, almost half of the predictors are dropped (average number of factors dropped is 44.64%), and the new model achieves similar performance.

For boundary tone, the accuracy of the rules learned from the new model is higher than the original model. For all other three models, the accuracy is slightly less but very close to the old models. Another interesting observation is that the pitch accent model appears to be more complicated than the other models. Twelve features are kept in this model, which include syntactic, semantic and intonational features. The other three models are associated with fewer features. The boundary tone model appears to be the simplest with only 4 features selected.

A similar experiment was done for data combined from the two speakers. An additional variable called "speaker" is added into the model. Again, the data is analyzed by the generalized linear model. The results show that "speaker" is consistently selected by the system as an important factor in all 4 models. This means that different speakers will result in different intonational models. As a result, we based our experiments on a single speaker instead of combining the data from both speakers into a single model. At this point, we carried out no other experiments to study speaker difference.

4.2.3 Sequential Rules

The simplified model acquired from Experiment 2 was quite helpful in reducing the complexity of the remaining experiments which were designed to take the intra-sentential context into consideration. Much of intonation is not only affected by features from isolated words, but also by words in context. For example, usually there are no adjacent intonational or intermediate phrase boundaries. Therefore, assigning one boundary affects when the next boundary can be assigned. In order to account for this type of interaction, we extract features of words within a window of size 2i+1 for i=0,1,2,3; thus, for each experiment, the features of the i previous adjacent words, the i following adjacent words and the current word are extracted. Only the salient predictors selected by experiment 2 are explored here.

The results in Table 3 show that intra-sentential context appears to be important in improving the performance of the intonation models. The accuracies of break index, phrase accent and boundary tone model, shown in the "Accuracy" columns, are around 90% after the window size is increased from 1 to 7. The accuracy of pitch accent model is around 80%. Except the boundary tone model, the best performance for all other three models improve significantly over the simple model with p=0.0017 for break index model, p=0 for both pitch accent and phrase accent model. Similarly, they are also significantly improved over the model without context information with p=0.0135 for break index, p=0 for both phrase accent and pitch accent.

4.3 The Rules Learned

In this section we describe some typical rules learned with relatively high accuracy. The following is a 5-word window pitch accent rule.

IF ACCENT1=NA and POS=adv
THEN ACCENT=H* (12/0)

This states that if the following word is deaccented and the current word's part of speech is "adv", then the current word should be accented. It covers 12 positive examples and no

Size	Break Index			Pitch Accent			Phrase Accent			Boundary tone		
	Accuracy	rule #	condition#	Accuracy	rule #	condition#	Accuracy	rule #	condition#	Accuracy	rule #	condition#
1	87.94%	7	18	73.87%	11	20	86.72%	5	15	97.36%	2	4
3	89.87%	5	11	78.87%	11	25	88.22%	7	15	97.36%	2	4
5	89.86%	8	26	80.30%	12	29	90.29%	8	23	97.15%	2	4
7	88.44%	8	20	77.73%	11	20	89.58%	9	26	97.07%	3	5

Table 3: System performance with different window size

negative examples in the training data.

A break index rule with a 5-word window is:

IF BB1=CB and SPPOS1=relative-pronoun
THEN INDEX=3 (23/0)

This rule tells us if the boundary before the next word is a clause boundary and the next word's semantic parent's part of speech is relative pronoun, then there is an intermediate phrase boundary after the current word. This rule is supported by 23 examples in the training data and contradicted by none.

Although the above 5-word window rules only involve words within a 3-word window, none of these rules reappears in the 3-word window rules. They are partially covered by other rules. For example, there is a similar pitch accent rule in the 3-word window model:

IF POS=adv THEN ACCENT=H* (22/5)

This indicates a strong interaction between rules learned before and after. Since RIPPER uses a local optimization strategy, the final results depend on the order of selecting classifiers. If the data set is large enough, this problem can be alleviated.

5 Generation Architecture

The final rules learned in Experiment 3 include intonation features as predictors. In order to make use of these rules, the following procedure is applied twice in our generation component. First, intonation is modeled with FUF/SURGE features only. Although this model is not as good as the final model, it still accounts for the majority of the success with more than 73% accuracy for all 4 intonation features. Then, after all words have been assigned an initial value, the final rules learned in Experiment 3 are applied and the refined results are used to generate an abstract intonation description represented in the Speech Integrating Markup Language(SIML) format (Pan and McKeown, 1997). This abstract description is then transformed into specific TTS control parameters.

Our current corpus is very small. Expanding the corpus with new sentences is necessary.

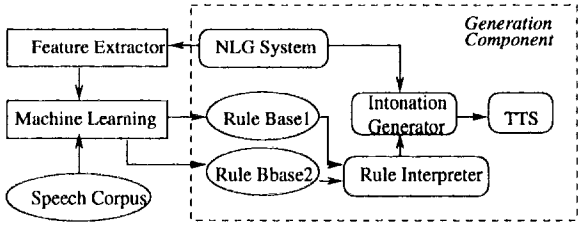

Figure 2: Generation System Architecture

Discourse, pragmatic and other semantic features will be added into our future intonation model. Therefore, the rules implemented in the generation component must be continuously upgraded. Implementing a fixed set of rules is undesirable. As a result, our current generation component shown in Figure 2 focuses on facilitating the updating of the intonation model. Two separate rule sets (with or without intonation features as predictors) are learned as before and stored in rulebase1 and rulebase2 respectively. A rule interpreter is designed to parse the rules in the rule bases. The interpreter extracts features and values encoded in the rules and passes them to the intonation generator. The features extracted from the FUF/SURGE are compared with the features from the rules. If all conditions of a rule match the features from FUF/SURGE, a word is assigned the classified value (the RHS of the rule). Otherwise, other rules are tried until it is assigned a value. The rules are tried one by one based on the order in which they are learned. After every word is tagged with all 4 intonation features, a converter transforms the abstract description into specific TTS control parameters.

6 Conclusion and Future Work

In this paper, we describe an effective way to automatically learn intonation rules. This work is unique and original in its use of linguistic features provided in a general purpose NLG tool to build intonation models. The machine-learned rules consistently performed well over all intonation features with accuracies around 90% for break index, phrase accent and boundary tone.

For pitch accent, the model accuracy is around 80%. This yields a significant improvement over the baseline models and compares well with other TTS evaluations. Since we used different data set than those used in previous TTS experiments, we cannot accurately quantify the difference in results, we plan to carry out experiments to evaluate CTS versus TTS performance using the same data set in the future. We also designed an intonation generation architecture for our spoken language generation component where the intonation generation module dynamically applies newly learned rules to facilitate the updating of the intonation model.

In the future, discourse and pragmatic information will be investigated based on the same methodology. We will collect a larger speech corpus to improve accuracy of the rules. Finally, an integrated spoken language generation system based on FUF/SURGE will be developed based on the results of this research.

7 Acknowledgement

Thanks to J. Hirschberg, D. Litman, J. Klavans, V. Hatzivassiloglou and J. Shaw for comments. This material is based upon work supported by the National Science Foundation under Grant No. IRI 9528998 and the Columbia University Center for Advanced Technology in High Performance Computing and Communications in Healthcare (funded by the New York state Science and Technology Foundation under Grant No. NYSSTF CAT 97013 SC1).

References

J. Bachenko and E. Fitzpatrick. 1990. A computational grammar of discourse-neutral prosodic phrasing in English. *Computational Linguistics*, 16(3):155–170.

Mary Beckman and Julia Hirschberg. 1994. The ToBI annotation conventions. Technical report, Ohio State University, Columbus.

L. Brieman, J.H. Friedman, R.A. Olshen, and C.J. Stone. 1984. *Classification and Regression Trees.* Wadsworth and Brooks, Monterrey, CA.

John Chambers and Trevor Hastie. 1992. *Statistical Models In S.* Wadsworth & Brooks/Cole Advanced Book & Software, Pacific Grove, California.

William Cohen. 1995. Fast effective rule induction. In *Proceedings of the 12th International Conference on Machine Learning.*

Mukesh Dalal, Steve Feiner, Kathy McKeown, Shimei Pan, Michelle Zhou, Tobias Hoellerer, James Shaw, Yong Feng, and Jeanne Fromer. 1996. Negotiation for automated generation of temporal multimedia presentations. In *Proceedings of ACM Multimedia 1996*, pages 55–64.

J. Davis and J. Hirschberg. 1988. Assigning intonational features in synthesized spoken discourse. In *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics*, pages 187–193, Buffalo, New York.

M. Elhadad. 1993. *Using Argumentation to Control Lexical Choice: A Functional Unification Implementation.* Ph.D. thesis, Columbia University.

Michael A. K. Halliday. 1985. *An Introduction to Functional Grammar.* Edward Arnold, London.

Julia Hirschberg. 1993. Pitch accent in context:predicting intonational prominence from text. *Artificial Intelligence*, 63:305–340.

Kathleen McKeown, Shimei Pan, James Shaw, Desmond Jordan, and Barry Allen. 1997. Language generation for multimedia healthcare briefings. In *Proc. of the Fifth ACL Conf. on ANLP*, pages 277–282.

Shimei Pan and Kathleen McKeown. 1997. Integrating language generation with speech synthesis in a concept to speech system. In *Proceedings of ACL/EACL'97 Concept to Speech Workshop*, Madrid, Spain.

Janet Pierrehumbert. 1980. *The Phonology and Phonetics of English Intonation.* Ph.D. thesis, Massachusetts Institute of Technology.

S. Prevost. 1995. *A Semantics of Contrast and Information Structure for Specifying Intonation in Spoken Language Generation.* Ph.D. thesis, University of Pennsylvania.

Jacques Robin. 1994. *Revision-Based Generation of Natural Language Summaries Providing Historical Background.* Ph.D. thesis, Columbia University.

Michelle Wang and Julia Hirschberg. 1992. Automatic classification of intonational phrase boundaries. *Computer Speech and Language*, 6:175–196.

S. Young and F. Fallside. 1979. Speech synthesis from concept: a method for speech output from information systems. *Journal of the Acoustical Society of America*, 66:685–695.