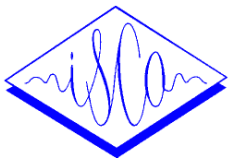

First International Conference on
Spoken Language Processing (ICSLP 90)
Kobe, Japan
November 18-22, 1990

Acoustic, Perceptual, and Linguistic Analyses of Intonation Contours in Human/Machine Dialogues[1]

Nancy A. Daly and Victor W. Zue

Spoken Language Systems Group
Laboratory for Computer Science
Massachusetts Institute of Technology
Cambridge, Massachusetts 02139 U.S.A.

ABSTRACT

This paper describes our research directed towards the quantification and use of prosodic cues in the intonation contours for different types of queries found in human/machine problem-solving dialogues. We ask three fundamental questions: First, what factors determine intonation encoding for queries? Second, how do these factors interact? Third, what are the implications for speech understanding? Our analysis is based on a corpus of spontaneous speech, containing several thousand sentences, collected in conjunction with the development of the MIT VOYAGER urban exploration and navigation system, under simulated human/machine dialogues. In our corpus, we found that over 90% of the WH-questions, such as *Where is MIT*, have low final boundary tones. For the YES-NO questions, such as *Is there a bank near Harvard*, on the other hand, only about 64% were found to have high final boundary tones. Our results, based on classification and regression tree analyses (CART), indicate that, while syntactic structure is the most important factor in predicting intonation contours, other factors such as the sentence's main verb and the speaker's sex are also important. We performed perceptual experiments in which subjects were asked to rate the appropriateness of a simple YES-NO answer on a 10-point scale. Our results confirm that listeners vary their judgments of YES-NO appropriateness based on factors other than final boundary tone.

INTRODUCTION

Prosody, the stress, rhythm and intonation of speech, is an important source of information for speech communication, spanning across many linguistic levels. Prosody encodes linguistic and extra-linguistic information in the speech signal, such as speaking rate, linguistic stress, syntax and semantics. While prosody has been regarded as important and useful in speech synthesis, there has been very little explicit use of prosodic information in speech recognition. This is partly due to a lack of understanding of variability in the encoding of prosodic information, particularly across many speakers.

This paper describes a study that addresses one aspect of prosodic encoding. We focus specifically on encoding of prosodic cues in intonation contours for different query types often found in human/machine problem-solving dialogues. We use speech elicited from many speakers during interactive dialogues with VOYAGER [1], an urban navigation and exploration speech understanding system under development at MIT.

Many linguists have suggested that there are two basic question types, those that are "test of fact" questions, best answered by *yes* or *no*, and those that are better answered in some other way [2,3]. YES-NO questions are characterized as typically ending with rising intonation, other questions with falling intonation. If these characterizations were true, such information would be sufficient for determining sentence type.

Our research has three fundamental goals. First, we attempt to discover what primary factors determine the encoding of prosodic information about query types. Second, we examine the interaction of these factors, and how they relate to the predictability of prosodic information. Finally, we explore the implications of our findings for speech synthesis and understanding.

We begin with a description of the design and transcription of the corpus used in our experiments. Next, we describe the acoustic, linguistic and perceptual experiments performed on the corpus. Finally, we discuss the possible use of our findings in the fields of speech synthesis and understanding.

CORPUS DESCRIPTION

The sentences that we used were drawn from a spontaneous speech corpus of human/machine dialogues collected under simulation, described in Soclof and Zue [4]. Data from forty-four male and forty-five female speakers that constituted the training and development sets were included, for a total of 4269 utterances. After examining part of the data, we adopted the convention of classifying the sentences into seven categories as shown in Table 1. Table 1 also includes the relative sizes of the categories. In this corpus, nearly 90% of the data are questions.

CATEGORY	%	EXAMPLE
Statement	6.6	*I am at MIT*
Command	1.6	*Show me Central Square*
Fragment	1.8	*Beacon Street*
Multiple	0.5	*I am at MIT how do I get to Steve's*
YES-NO Ques.	19.0	*Does Toscanini's serve ice cream*
WH- Ques.	70.2	*Where is the Royal East*
Clarification Ques.	0.4	*How about chinese food*

Table 1: Corpus utterance categories

ANALYSES AND RESULTS

Acoustic Labelling

To facilitate subsequent analyses, the intonation contours were first labelled. Rather than providing a detailed prosodic tran-

[1]This research was supported by DARPA under Contract N00014-89-J-1332, monitored through the Office of Naval Research.

scription, we have thus far only transcribed the end of each sentence. We adopted the scheme proposed by Pierrehumbert [5], in which all clauses in American English end with either a relatively high or low boundary tone. We postulate that perceived rising intonation corresponds to a high final boundary tone, and perceived falling intonation to a low final boundary tone.

Several acoustic phoneticians transcribed each of the utterances as LOW, HIGH or UNCERTAIN by listening to them. Each utterance was transcribed by at least two people, and the inter-transcriber agreement was over 90%. In cases of disagreements, additional transcribers were consulted to resolve conflicts.

Figure 1 shows the percentage of utterances with low final boundary tones for each of seven categories. This figure reveals that the final boundary tone is usually low for most categories.

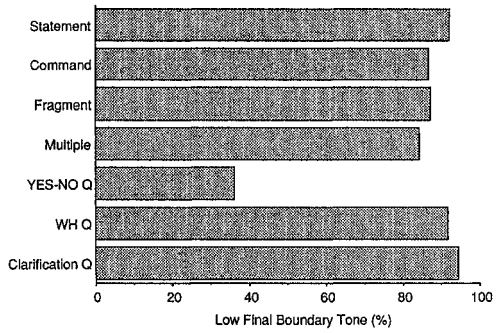

Figure 1: Percentage, by utterance category, of utterances marked as having low final boundary tones.

Factor Analysis

We see from Figure 1 that YES-NO questions are indeed accompanied by high final boundary tones. However, this trend is far from unequivocal; a full 36% have low final boundary tones. Clearly, the simplistic categorization described earlier cannot adequately account for our data. Other factors may play a role in the assignment of final boundary tones for so-called YES-NO questions. Pilot analyses we performed suggested our factor analysis may require that we first obtain a more detailed linguistic transcription, rather than a simple YES-NO/NOT YES-NO dichotomy.

Linguistic transcription was accomplished by passing each sentence through TINA, a probabilistic natural language system developed in our group [6]. From the orthographic transcription of a sentence, TINA produces a parse tree, incorporating syntactic and semantic constraints[2]. Figure 2 shows an example of a parse tree produced by TINA. While any of the labels of TINA's parse nodes can be used for factor analysis, we have restricted ourselves to include only the first level. In this example, only Q-SUBJECT (i.e., query subject) and BE-QUESTION (i.e., sentences with verb "to be" as the main verb) are used. Of the 3825 questions in the corpus, 3014 (78.8%) can be parsed by TINA and have unequivocal boundary tone assignments; none were clarification questions. Therefore, our factor analysis is restricted to only these sentences. In addition to the top level parse node la-

[2]The orthographic transcriptions of these spontaneous sentences often include hesitations, false starts, and filled pauses. These spontaneous speech events were removed from the transcription before natural language processing.

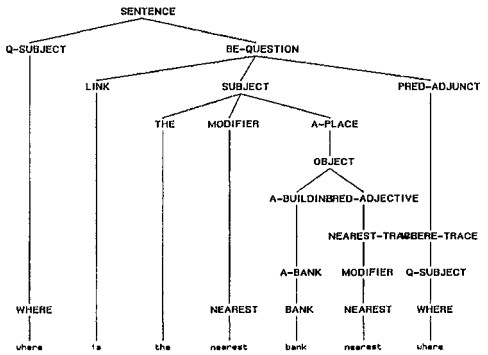

Figure 2: TINA parse tree for the sentence *Where is the nearest bank.*

bels produced by TINA, we included several other factors such as the main (conjugated) verb of the sentence and sex of the speaker. Discourse level information about similarity of the linguistic structure to the preceding sentence was also included. We noticed in our preliminary analysis that, when two utterances were structurally similar, speakers sometimes modified intonation contours in order to draw the listener's attention to the contrasting portion of the sentences. The specific factor used in our analysis was a tag that indicated whether or not a given utterance had the same first level parse as its predecessor in a given dialogue.

The importance of the various factors was established using CART, a classification and regression tree analysis scheme [7]. The available data were divided into a training set (2295 tokens from sixty-nine speakers) and test set (719 tokens from the remaining twenty speakers). Half of the training data was used to develop a classification tree using the final boundary tone value as the response variable and the factors described above as the explanatory variables. The remaining half of the training data was set aside for cross-validation. A categorical split was performed for each node, based on the optimal (smallest) split impurity yielded. Splitting continued until either the node was pure or all explanatory variables had been exhausted, resulting in seventy-one terminal nodes. The cross-validation data were then passed through the same tree, and the tree was recursively pruned until both the total number of misclassifications and node impurities were reduced to the optimal level. The resulting (pruned) tree contained five nodes, as shown in Figure 3. The tree was evaluated for robustness on the test set. We obtained classification accuracy rates of 87.5%, 88.6% and 86.2% on the development, cross-validation and test sets, respectively, when each set was passed through the pruned tree.

Figure 3 shows that the most important factor for determining whether a sentence should have high or low final boundary tone is indeed the question type, i.e., YES-NO or WH-. Over 90% of the WH- questions in the test set have low final boundary tones. For YES-NO questions, the presence of a high final boundary tone seemed to depend on the main verb. For sentences such as *Can*

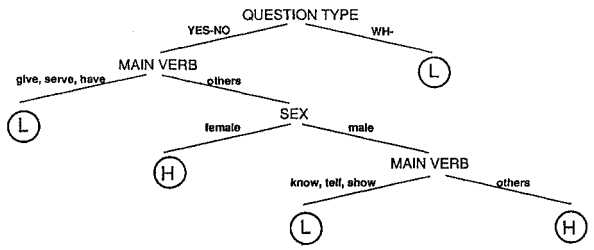

Figure 3: Classification tree showing predictability of final boundary tone for test data based on four parameters.

you give me directions to MIT, or *Can you show me the nearest restaurant*, low final boundary tones were often observed. Speaker sex also seemed to play a role, with female speakers much more likely to have low final boundary tones than male speakers.

Perceptual Analysis

The above analysis suggests that, at least for our data, a direct association of YES-NO questions with high final boundary tones may not be appropriate. It is possible that some of the so-called YES-NO questions, e.g., *Do you know what time it is*, actually require more than a simple *yes* or *no* answer. To explore this issue further, two perceptual experiments were conducted. In the first experiment, subjects were given *written* questions and asked to judge the appropriateness of a simple *yes* or *no* answer on a scale of 1 (i.e., entirely appropriate) to 10 (i.e., entirely inappropriate). In the second experiment, subjects *listened* to the same questions and were asked to evaluate them in the same way. The data consisted of 451 questions randomly selected from our corpus; 102 (22.6%) of them were WH- questions that served as controls. Five subjects participated in the experiments; no subjects heard the same stimuli.

Means and standard deviations were calculated for the two types of questions across subjects, and the results are shown in Table 2. Subjects virtually always considered WH- questions to be inappropriately answered by a simple *yes* or *no*. However, labels for YES-NO questions varied widely. There is also a statistically significant difference ($p = 0.001$) between the means of the reading and listening experiments. This result suggests that subjects answered differently when acoustic information, presumably the intonation contour, was available to them. But what specific aspect of this acoustic information was responsible for the shift? To explore this issue further, we examined the pairs of reading/listening responses as a function of final boundary tone. Specifically, we examined the distribution of the *difference* in the subjects' responses between the reading and listening experiments. The results indicate that there is a significant difference in responses for sentences having low final boundary tones. That is, a subject is more likely to lower his or her ranking when a low final boundary tone is present. The results shown in Table 2 suggest that listeners' responses varied greatly for YES-NO questions. One may wonder whether this variation is correlated with the final boundary tone. Table 3 shows listeners' answers as a function of final boundary tone. We arbitrarily collapsed the 10-point scale into three bins: NOT YES-NO (1-3), UNCERTAIN (4-7), and YES-NO (8-10). Nearly half the sentences with high final boundary tones were perceived by listeners to be appropriate YES-NO questions, whereas 45% of those with low final boundary tones were perceived to be inappropriate.

Test	μ_w	σ_w	N_w	μ_y	σ_y	N_y
Reading	1.00	0	102	6.24	3.50	349
Listening	1.01	0.10	102	5.27	3.40	349

Table 2: Means and variances of subject responses for both perceptual experiments. Statistics for WH- questions have w subscripts, while those for YES-NO questions have y subscripts.

FBT	NOT YES-NO	UNCERTAIN	YES-NO	Total
HIGH	25.6%	25.6%	48.8%	160
LOW	45.0%	30.2%	24.9%	189

Table 3: Categorization of listener responses by final boundary tone (FBT). The 10-point scale used by listeners was collapsed into three classes.

Our results seem to indicate that listeners use the final boundary tone as an additional source of information in deciding whether or not a question is YES-NO. Our earlier factor analysis has resulted in a classification tree, in which factors such as the main verb and the speaker's sex were found to be important in final boundary tone assignments. To see how the perceptual results might be reconciled with the factor analysis, sentences used in the listening tests were passed through the tree and assigned appropriate HIGH or LOW labels. As Table 4 indicates, there is a significant difference in listeners' scores assigned to HIGH and LOW tokens. We conclude that the factor analysis and perceptual experiment results are mutually supportive.

FBT	μ	σ	N
HIGH	6.26	3.04	257
LOW	1.71	2.04	194

Table 4: Means and variances of listeners' responses when tokens are relabelled based on their positions in the classification tree in Figure 3.

DISCUSSION

This paper describes our study of the use of prosodic cues in intonation contours for different types of queries found in human/machine problem-solving dialogues. While the data that we used were drawn from a particular spoken language system, namely VOYAGER, we believe that our findings can be generalized to other tasks involving database queries.

Our results suggest that, while WH- questions are invariably accompanied by low final boundary tones, prosodic encoding of the so-called YES-NO questions is somewhat more complicated. High final boundary tone assignment for these questions depends on a variety of factors, including the main verb of the sentence and the speaker's sex. This finding seems to be supported by results from perceptual experiments. While the main verb appears to be

an important factor, our results may be interpreted in a slightly different light as well. Closer examination of sentences containing main verbs such as "give" (e.g., *Can you give me directions to MIT?*) or "know" (e.g., *Do you know the telephone number of the Hyatt Regency?*) suggests that these sentences may be considered indirect speech acts [8]. They can be interpreted as polite forms of requests, thus the resulting low final boundary tones.

Information encoded in final boundary tones for questions is relevant to several areas of research in spoken language systems. For speech synthesis, a greater understanding of the relationship between final boundary tone values and the linguistic structures of a sentence can aid in improving the naturalness of synthesized speech. For speech recognition, this information can provide an independent knowledge source for disambiguating sentence syntax and speaker's intention [8]. In addition, prosodic information derived from the speech signal, such as final boundary tone values, can be helpful for natural language processing. For example, a probabilistic parser like TINA [9] can make use of this information, in addition to probabilities derived from word class information, for the assignment of the probabilities of partial theories. This would increase computational efficiency, as it would proceed with the correct parse before attempting other less likely hypotheses. Finally, our perceptual experiment results may be applied to improving response generation in human/machine dialogues. Interactions with machines are facilitated for the user if a machine responds to a query in the manner he or she expects.

Since the practical use of final boundary tone information depends on its measurability, it is necessary to find its salient acoustic correlates. In a pilot study, we have found that the final boundary tone occurs during the last syllable of an utterance, and that the fundamental frequency may contain sufficient information to determine its value. Thus our search for the acoustic correlates of the final boundary tone was restricted to the final syllable, and only the fundamental frequency was used.

Using a phonetically transcribed subset of the corpus, we conducted an acoustic analysis to test our proposal. This subset consisted of 1288 utterances (229 HIGH and 1059 LOW) collected from sixteen male and twelve female speakers. Several different measurements of fundamental frequency were made, including its averages over the utterance and the last syllable nucleus, and the change in fundamental frequency within the last syllable nucleus. Two-thirds of the data (857 utterances) was used to determine the best measurement for distinguishing high final boundary tone from low final boundary tone utterances, and the remaining one-third was used to test the robustness of the measurement. Training and testing were performed for male and female speakers both separately and combined. In all cases, the most robust measurement for separating high from low final boundary tone utterances was the difference between the average fundamental frequency in the last syllable nucleus and the average fundamental frequency over the entire utterance. The results of these experiments are shown in Table 5. The final boundary tone is more accurately measured by the proposed correlate in female speech than in male. This may be due to two factors. First, the variability of female pitch is greater than that of males, meaning that final boundary tones in the former are marked by greater pitch excursions. Secondly, it is difficult to accurately track pitch when speech contains vocal fry, and this phenomenon is far more common in male speakers than in females.

Data Set	Training (%)	Testing (%)
Male	86.4	87.9
Female	93.7	92.6
Combined	90.2	89.8

Table 5: Accuracy of proposed acoustic correlate for final boundary tone measurement (difference between average fundamental frequency of the last syllable nucleus and average fundamental frequency over the utterance).

As a final note, we would like to point out that prosodic encoding is much more salient in spontaneously produced speech than read speech. The corpus we used consisted of read/spontaneous utterance pairs, and we have observed that read versions are usually accompanied by declinations in fundamental frequency contours, regardless of sentence type. Such anomalies associated with read speech greatly obscure acoustic information, making it difficult to discover the underlying prosodic encoding. We strongly recommend that future prosodic analyses be carried out using spontaneous speech, in which speakers are actively participating in speech communication.

REFERENCES

[1] Zue, V., J. Glass, D. Goodine, H. Leung, M. Phillips, J. Polifroni, and S. Seneff, "The VOYAGER Speech Understanding System: Preliminary Development and Evaluation," *Proceedings, International Conference on Acoustics, Speech and Signal Processing*, April 1990.

[2] Harris, Z., "The Interrogative in a Syntactic Framework," from *Questions*, edited by Henry Hiz, Dordrecht: D. Reidel Publishing Company, 1978, pp 1-36.

[3] Hintikka, J., *The Semantics of Questions and the Questions of Semantics: Case Studies in the Interrelations of Logic, Semantics, and Syntax*. Amsterdam: North-Holland Publishing Company, 1976.

[4] Soclof, M. and V. Zue, "Collection and Analysis of Spontaneous and Read Corpora for Spoken Language System Development," these *Proceedings*.

[5] Pierrehumbert, J., "The Phonology and Phonetics of English Intonation," PhD Thesis, Massachusetts Institute of Technology, 1980.

[6] Seneff, S., "TINA: A Probabilistic Syntactic Parser for Speech Understanding Systems," *Proceedings, DARPA Speech and Natural Language Workshop*, February 1989.

[7] Breiman, L., J. Friedman, R. Olshen, and C. Stone, *Classification and Regression Trees*, Monterey: Wadsworth and Brooks/Cole Advanced Books and Software, 1984.

[8] Hirschberg, J., "Distinguishing Questions by Contour in Speech Recognition Tasks," *Proceedings, DARPA Speech and Natural Language Workshop*, October 1989, pp 22-34.

[9] Seneff, S., "Probabilistic Parsing for Spoken Language Applications," *Proceedings, International Workshop in Parsing Technologies*, August 1989.