Distinguishing Questions by Contour in Speech Recognition Tasks

Julia Hirschberg
AT&T Bell Laboratories
600 Mountain Avenue
Murray Hill NJ 07974

October 27, 1989

1 Can We Predict Intonation?

It is generally acknowledged today that, while the intonational features speakers select when they utter a sentence are not *determined* by the syntax, semantics or discourse context of that sentence, knowledge of these factors can help to constrain the possible intonational features speakers are likely to choose. So, while intonational variation poses a challenge to speech recognition in one sense – in presenting yet another indicator of over-all utterance meaning to be recognized – regularities noted between intonational features and the syntax, semantics and discourse features of an utterance also present rich possibilities for help in the recognition task.

The many-to-many mapping between intonational features and syntactic and discourse features can be illustrated by considering the various ways of uttering the sentences in (1).[1]

(1) a. 560 CAN KIRK GET TO KODIAK BY MONDAY

b. Kirk can get to Kodiak by Monday.

For example, a senior officer might choose a falling pitch contour over (1a) to convey an indirect request that *Kirk* reach *Kodiak* by Monday. A less senior speaker, however, might produce (1a) with rising (yes-no question) intonation, conveying merely a request for information. Alternatively, the syntactically related but distinct form of (1a), (1b), might be produced with rising intonation to convey the same request for information, or with falling intonation to convey the information thus requested. So, different contours can be used over the same

[1]Example sentences in this paper are taken from the DARPA Resource Management database.

sentence to convey different meanings, and the same contour may be used over sentences differing only in syntactic structure to convey a similar meaning.

Despite such possibilities for variation, research in intonational meaning and more practical application of such research in speech synthesis indicates that there *are* regularities that recognition systems may be able to utilize. For example, knowledge of likely relationships between syntax and intonation tell us that, when (1b) is uttered in natural speech, it will be more likely to be said with falling intonation than with rising. And given that we know the speaker of (1a) and have some rough idea of that speaker's authority, we can also predict whether that speaker will be likely to use rising intonation or not, based on whether that speaker will be likely to be trying to convey an indirect request or simply to gain information. So, knowledge of the structure of an utterance and knowledge of the overall context in which it is uttered help to constrain the set of intonational possibilities. Thus, intonational features such as contour type provide indirect evidence as to what the syntactic structure of the associated sentence might be, given that we know the likelihood that a sentence like (1a) might be uttered to convey an indirect request rather than a request for information. That is, in the general case, given that we know the likely utterer of (1a) to be a clerk, we will expect (1a) to be uttered as a request for information, and, thus with rising intonation.

2 Can We Use Intonational Information to Aid Speech Recognition?

Interest in using higher-level intonational information such as pitch contour, intonational phrasing, and pitch accent placement to aid speech recognition has been intermittent.[Lea79, Pie83, Wai88] Progress in this area has been hindered by a) the difficulty of extracting higher level intonational characteristics automatically with any reliability; b) the lack of representations of the features to be extracted such that information can be incorporated into the recognition process; and c) an imperfect understanding of the particular constraints syntax, semantics and discourse features impose on a speaker's choice of intonational features. Thus, practical problems of feature detection have gone hand in hand with more theoretical issues of representation and interpretation. However, there has been some progress in developing algorithms to extract and identify at least partial information about higher-level intonational features, such as differentiation of stressed and unstressed syllables and distinction of rising from falling contours.

At this stage, it does seem likely, that particular recognition tasks and particular domains will find some higher-level intonational cues more useful than others. For testing the utility of predicting the syntactic 'type' of an utterance from its intonational contour, for example, domains in which there are broad

classes of utterances which can be reliably partitioned according to both intonational and syntactic category appear promising. Database query tasks, for example, where there is a reasonable balance between inverted yes-no questions[2], which are commonly uttered with final rising intonation, and *wh*-questions[3], or imperatives[4], which are both commonly uttered with final fall — and in which there is relatively little likelihood of speech act ambiguity, seem well-suited to such an experiment. The DARPA Resource Management (*RM*) task thus seemed a good place to look for such distinctions.

In domains such as this, we might expect that distinguishing likely yes-no questions from other sentences might be a useful augmentation for traditional recognition methodologies, acting as a filter on matches proposed by the recognizer or even providing an initial state in a regular grammar partitioned by broad syntactic 'type'.[5] The utility of adding such information is supported by certain classes of recognition errors, such as those illustrated in (2).[6] These errors represent instances in which the ability to distinguish yes-no questions intonationally from *wh*-questions, imperatives, and other sentence 'types' typically uttered with falling intonation might serve as an aid to recognition (In each case, the (a) sentences represents the test sentence and the (b) sentence represents the recognizer's hypothesis.):

(2) a. REF: IS kennedy+s arrival hour in pearl harbor AFTER ** fifteen hundred hours
HYP: GIVE kennedy+s arrival hour in pearl harbor HAVE TO fifteen hundred hours

b. REF: WHAT IS the total fuel aboard THE mars
HYP: WAS ** the total fuel aboard *** mars

c. REF: IS shasta within six kilometers of thirteen north forty east
HYP: THE shasta within six kilometers of thirteen north forty east

d. REF: WHEN+LL enterprise next be in home PORT
HYP: WILL enterprise next be in home PORTS

e. REF: *** FIND speeds available for england and fox
HYP: ARE THE speeds available for england and fox

That is, the test sentence represents a sentence type likely to be uttered with an intonational contour which would distinguish it from the sentence incorrectly

[2]Sentences in which aux- or copula-inversion has occurred, such as *'IS MARS+S LAST LAT IN NORTH ATLANTIC OCEAN '* where the copula *is* has been inverted (cf. *'Mars's last lat is in North Atlantic Ocean.'*.

[3]Questions beginning with *who, what, when, where* or *how.*

[4]Such as *'DISPLAY METEOR+S LON USING OVERLAY BOX '*.

[5]That is, inverted yes-no questions might be separated from other syntactic constructions in the grammar.

[6]These were some of the errors made on the DARPA February89 training set by one of the Bell Labs recognizers.[Lee89, LRPW89]

hypothesized by the recognizer. Among these errors, distinguishing between '*when'll*' and '*will*' and between '*what is*' and '*was*' would appear to be particularly difficult tasks for a recognizer on acoustic grounds. In fact, about 8% of sentence errors made in this test were due at least in part to one of these two confusions. Table 1 shows all sentence errors in the test run in which yes-no questions were confused with *wh*-questions, imperatives, or declarative sentences.[7] (Column 2 shows the category of the actual utterance; column 3 show the category of the utterance recognized (yes-no question (ynq),*wh*-question (whq) or imperative (imp)); and column 4 show the lexical items confused.)

Table 1: 'Type' Errors on the DARPA Feburary '89 Test Set

Sentence Number	Type of Sentence	Type of Hypothesis	Items Confused
3	wh	ynq	how soon ⇒ has the
7	ynq	imp	is ⇒ give
13	whq	ynq	when'll ⇒ will
15	imp	ynq	clear ⇒ did
56	ynq	imp	is ⇒ give
61	whq	ynq	what is ⇒ was
68	whq	ynq	what is ⇒ was
86	whq	ynq	what's ⇒ was
104	ynq	decl	is ⇒ the
241	whq	ynq	what is ⇒ was
242	whq	ynq	when'll ⇒ will
247	imp	ynq	find ⇒ are
267	whq	ynq	what is ⇒ was
272	whq	ynq	what is ⇒ was
287	whq	ynq	what is ⇒ was
292	whq	ynq	what is ⇒ was

Total sentences incorrect: 128
Total sentence type errors: 16

Of the 16 errors which type of contour *might* have been able to prevent – on the assumption that yes-no questions should have been produced with

[7]Note of course that some of the mistaken hypotheses were not in fact grammatical, such as c) and (2d) above, so the assignment of sentence 'type' was based upon possible completions the longest initial grammatical string. So, '*WAS ** the date and hour of arrival in port r arkansas* ' was considered structurally a yes-no question.

rising intonation and other utterance types with falling intonation — 15 of the misrecognized utterances in fact were spoken with the 'likely' contour for their syntactic type. That is, in fifteen cases a yes-no question uttered with rising intonation was misrecognized as a syntactic type (*wh*-question or imperative) which would have been unlikely to have been uttered with rising intonation – or a non-yes-no question uttered with falling intonation was misrecognized as a yes-no question.

However, while these errors might thus have been filtered by this simple association between contour and sentence type, it is not at all clear how well this solution might generalize even to other sentences within the same domain. While yes-no questions are typically uttered with rising intonation in natural speech — and *wh*-questions and imperatives commonly uttered with utterance-final fall, it is not clear whether such distinctions appear with the same likelihood in sentences read in isolation, the data which most recognizers train and test upon. To investigate the possibility then of predicting structural distinctions from intonational ones, it is useful to examine the prosody of the training and test data itself.

3 Are Yes-No Questions Intonationally Distinguished in the *RM* Database?

To assess the potential for using contour to distinguish inverted yes-no questions from other constructions in current recognition tasks, I sampled inverted yes-no questions and *wh*-questions from the training and test data of the speaker independent *RM* database.[PFBP88] Of the 2810 sentence types in the *RM*

Table 2: Sentence Types and Tokens in the *RM* Database

Total S-types	2810 (100%)
Total Questions	≈ 1694 (60%)
Total YNQs	≈ 670 (24%)
Total WH-qs	≈ 1024 (36%)
Sample	
YNQs_type	50
YNQs_token	100
WH-qs_type	53
WH-qs_token	100

database, approximately 60% can be classed either as inverted yes-no questions (24%) or *wh*-questions (36%). I sampled 100 utterances of yes-no questions (from 50 types) and 100 utterances of *wh*-questions (from 53 types) to deter-

mine whether sentences were uttered with rising intonation or not.[8] The yes-no questions chosen were inverted copula questions in the present tense of the form '*Is* .... '; no alternative questions were included in the sample, since these tend to be uttered with falling intonation. Both sample yes-no questions and sample *wh*-questions were selected from among the sentences in the database fewer than 9 words in length, to minimize the likelihood of multiple intonational phrases in the utterance or of performance error in the production of the utterance.

Of the 100 yes-no questions sampled from the *RM* database, only 55 were uttered with final rise. Only 9 of the *wh*-questions were similarly uttered, with the majority uttered with falling or level pitch. While the latter results seem consistent with previous observation about the tendency of *wh*-questions to fall, the findings for yes-no questions seem far too low.

Table 3: *RM* Sample

Question Type	Non-Rising	Rising	Total
YNQ	45	55	100
WH-Q	91	9	100

To test the representativeness of the contours in the sample, I examined samples of 50 inverted yes-no questions and 50 *wh*-questions from the TIMIT database (All were of distinct types.).[9] The results, presented in Table 4 appear much more consistent with observations of questions asked in natural speech. Thus, while only 55% of yes-no questions in the *RM* database were uttered

Table 4: TIMIT1 Sample

Question Type	Non-Rising	Rising	Total
YNQ	9	41	50
WH-Q	46	4	50

with rising intonation, over 80% of yes-no questions in the TIMIT1 sample rose. Production of *wh*-questions appears similar in both databases, with only 8% of the TIMIT1 *wh*-questions and 9% of the *RM* *wh*-questions uttered with final rise.

[8]Note that this distinction oversimplifies the distinctions observed in natural speech between question-rise and question-rise, but the results of this simple analysis did not warrant more refinement.

[9]TIMIT has a much lower proportion of questions than the *RM* database, with only 142 questions among TIMIT1's 1726 sentence types, some of them not syntactic yes-no questions or *wh*-questions. These were not considered in the sample.

The question then arises: why are yes-no questions produced so differently in the *RM* sentences than in the TIMIT1 sentences? Several explanations come to mind. First, one might hypothesize that certain yes-no questions would be more likely than others to be uttered with falling intonation, depending upon their semantic content. Those that might be interpreted as indirect requests, for example, like (1a), might tend to be uttered with falling intonation, as noted above. If the *RM* sentences were ambiguous with respect to speech act, then readers might favor falling intonation with such sentences. Unfortunately, the contours in Figure 1 illustrate a not-uncommon finding in the sample of yes-no questions selected from the *RM* database – a pair of utterances of the same yes-no question, one uttered with a rising contour and the other with a falling contour; 19 other pairs reflect this dichotomy. The commonality of varying contours over the same sentence type together with the fact that I avoided yes-no questions with seeming potential for speech act ambiguity in this domain makes this explanation unlikely.

A similar alternation is evident among the (many fewer) *wh*-questions uttered with rising intonation, which are illustrated by f0 contours from the same two speakers from Figure 1. For the *wh*-questions, it appears likely that speaker variability might account for the rising contours, since 6 of the 9 rising *wh*-questions were produced by a single speaker. However, a similar account does not appear possible for the 45 falling yes-no questions; no single speaker was responsible for more than 3.

Another explanation for the results of Table 3 is suggested by the sort of contour illustrated in Figure 3. The lack of variation in pitch prominence and range shown in these f0 contours seems to be not atypical of much of the *RM* database — and appears to reflect a lack of engagement in the subjects, an absence of any attempt to reflect sentence 'meaning' in their productions, and — in the type of disfluencies that mark much of the data — some difficulty in performing the task. In short, the data do not appear to have been produced so as to maximize their reflection of the higher-level intonational characteristics of natural speech.

So far the discussion of contour variation in the *RM* database has focussed on the disparity between subject performance in these sentences and observations from natural speech. Nonetheless, even with data not intended to provide material for detecting and using contour variation, the possibilities for using such intonational cues are still substantially supported. The difference between subject tendency to use rising intonation with *wh*-questions and with yes-no questions, like the difference in propensity to use non-rising intonation with yes-no questions and *wh*-questions, is in fact still quite significant.[10] And, while yes-no questions in the *RM* database are only uttered with rising intonation 55% of the time, note that approximately 86% of utterances with rising intonation in

[10]For example, a simple χ^2 test of the data in Table 3 is significant at the .001 level (The χ^2 statistic is 48.6 with df=1.)

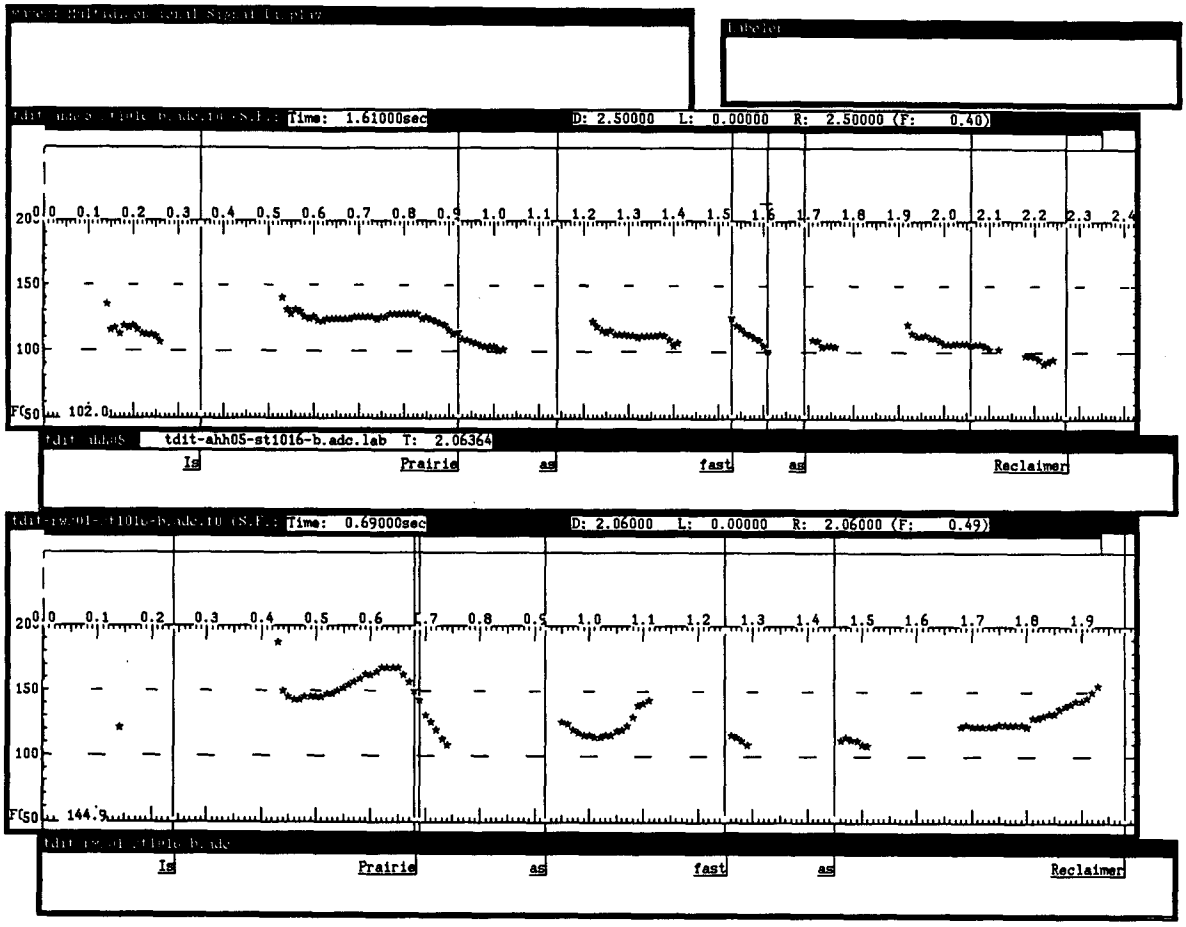

Figure 1: Rising and Falling Contours over the Same Yes-No Question

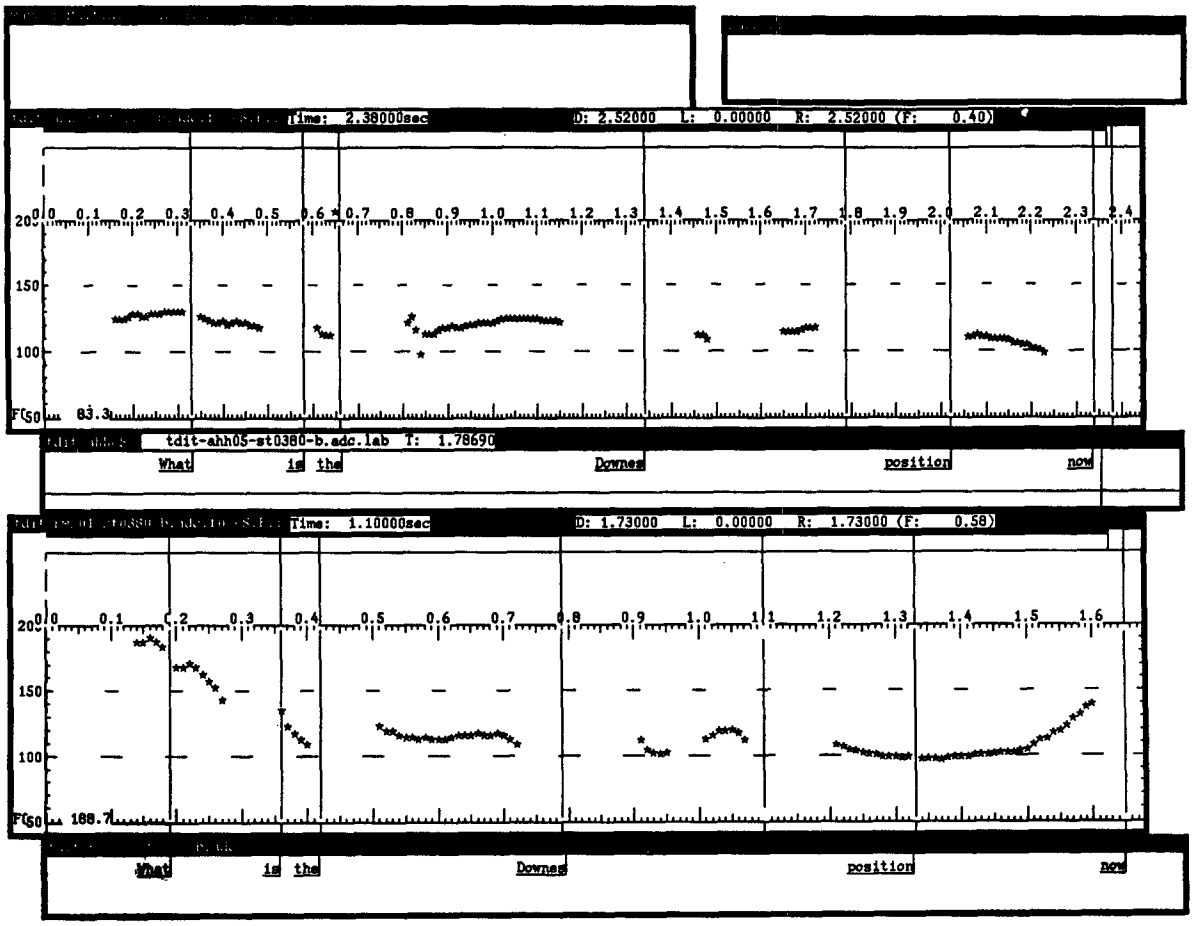

Figure 2: Rising and Falling Contours over the Same WH-Question

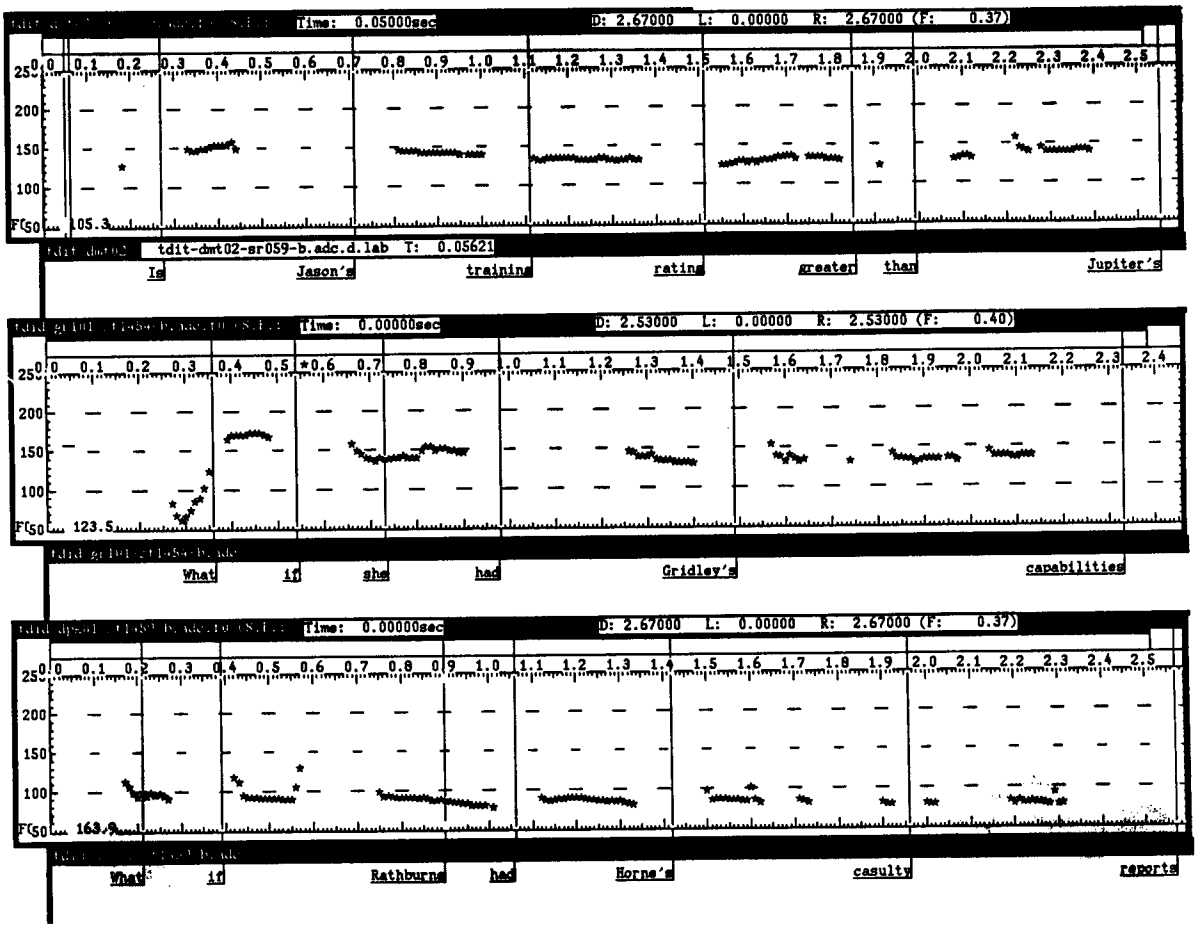

Figure 3: Yes-No questions and WH-Questions

the sample are indeed yes-no questions. And utterances characterized by falling intonation are fully twice as likely to be *wh*-questions as yes-no questions, according to our sample. So, even though we might expect a contour distinction to be even more successful in recognizing data from, say, TIMIT1, even the apparently much less 'natural' data of the *RM* sentences provides a good case for the idea that sentence classes might indeed be distinguished by the contour with which they are uttered.

4 Conclusion

While the *RM* sentences thus appear less than ideal in providing data for exploring the notion that distinguishing among general classes of pitch contour can be useful in distinguishing among structural classes of sentence for speech recogniz ers, this database nonetheless provides evidence that even subjects reading sentences in isolation will approximate some of the distinctions made in real speech. As one example, the association between contour type and sentence type appears significant enough to permit overall contour type to serve as a filter for recognition – at least for rising contours. That is, a rising contour should be a fairly reliable indicator of a yes-no question.

Nonetheless, it is also clear that this association should be providing even better discriminatory power than it does in this database. If future data collection efforts are to support more sophisticated uses of higher-level intonational information in the aid of speech recognition, then the standard data-collection paradigm of sentences read in isolation must certainly be abandoned. Just as recognizing connected speech poses different problems from isolated word recognition, recognizing real, interactive speech poses different problems from recognizing isolated sentences. In natural speech, speakers use intonation to convey the meaning of a sentence and to convey relationships between that meaning and the meanings of other sentences. But speakers will not use prosody to convey meaning unless they understand the meaning to be conveyed. And speakers will not use prosody to convey relationships among sentences in a discourse if they are not generating larger pieces of text. So long as recognition systems are tested merely on isolated sentences, of course, the difference between training and test data will be less important. But systems that are expected to supported even minimally longer dialogues will suffer, since intonational contours, phrasing, and stress assignment in interactive speech will vary significantly from isolated sentence data. In sum, training data and test data should mimic as much as possible the speech recognizers hope to recognize if both the problems presented by intonational variability and the possiblities presented by intonational regularities are to be adequately explored.

References

[Lea79] W. A. Lea. Prosodic aids to speech recognition. In W. A. Lea, editor, *Trends in Speech Recognition*, pages 166–205. Prentice-Hall, Englewood Cliffs NJ, 1979.

[Lee89] Chin-Hui Lee. Personal Communication, 1989.

[LRPW89] Chin-Hui Lee, Lawrence R. Rabiner, Roberto Pieraccini, and Jay G. Wilpon. Acoustic modeling of subword units for large vocabulary speaker independent speech recognition. In *Proceedings*. DARPA Speech and Natural Language Workshop, October 1989.

[PFBP88] P. Price, W. M. Fisher, J. Bernstein, and D. S. Pallett. The DARPA 1000-word Resource Management Database for continuous speech recognition. In *Proceedings*, volume 1, pages 651–654, New York, 1988. ICASSP88.

[Pie83] Janet B. Pierrehumbert. Automatic recognition of intonation patterns. In *Proceedings*, pages 85–90, Cambridge MA, 1983. Association for Computational Linguistics.

[Wai88] Alex Waibel. *Prosody and Speech Recognition*. Pitman Publishing, London, 1988.