# PROSODIC CUES TO WORD USAGE

Karen Ward and David G. Novick

Center for Spoken Language Understanding
Oregon Graduate Institute of Science & Technology
20000 N.W. Walker Road
Portland, Oregon 97006
USA

wardk@cse.ogi.edu, novick@cse.ogi.edu

## ABSTRACT

In this study we examined prosodic characteristics of a word used in several distinct senses in a task-oriented corpus of spontaneous speech. We compared the pitch characteristics of the word "right" used in three different senses: as an acknowledgment, as a direction, and as an affirmative answer to a question. Significant differences in intonation for different classes of usage were found, although the differences are not reliable enough to allow systems to use prosody alone to distinguish between usages. These results suggest that pitch change as reported by a pitch tracker could serve as a confirming cue when analyzing ambiguous speech recognizer output, or could serve as input to a probabilistic parser to aid in disambiguating senses of homonyms.

## 1. INTRODUCTION

To improve the robustness and performance of spoken language understanding systems, we would like our systems to incorporate computational models of a variety of conversational cues [1]. Before we can take full advantage of these new resources, we must first better understand how various spontaneous speech phenomena interrelate. We need to understand how cues from various sources contribute to or eliminate possible interpretations of an utterance, and how these cues might be usefully exploited in spoken language understanding systems.

In this study, we examine a potential interrelationship between word prosody and word usage. Word usage and word sense disambiguation are major concerns in text understanding systems, but in spoken language understanding system these concerns are minimized by sharply constraining the vocabulary and by imposing a strong task orientation on a word's interpretation. Often only one word usage is considered for each item in the vocabulary. As vocabularies for these systems expand and become more general, however, this technique will not be adequate.

The natural language processing community has made great strides in word sense disambiguation. McRoy [2] discusses the many cues that have proved valuable in textual word sense discrimination: morphology, syntactic tag, domain context, collocation (common expression), semantic context, syntax, role, discourse focus. While these are expected to be important in future spoken language understanding systems, they are unlikely to be adequate alone because of the several important differences between text and speech. First, the speech recognizer may have made errors, so the words that the system is trying to analyze may be incorrect. Also, spoken language lacks the visual punctuation cues of text; it may be difficult even to determine whether the speaker has finished the utterance. The utterance itself is often full of false starts and ungrammaticalities not normally encountered in text. Finally, extensive syntactic and semantic processing can be expensive and slow in the context of a system expected to participate in a real-time conversation.

In understanding spontaneous speech, a system will need to make use of as many cues as possible to compensate for the great variability in spoken language input. Where possible, we would like to drawn on low-level cues that are simple and fast to process, so that slower and more complex analysis can be reserved for those inputs that require it.

Phrasal tunes offer valuable cues to the speaker's intentions. Pierrehumbert and Hirschberg [3] proposed that tunes signal relationships between the propositional content and the mutual beliefs of the participants. More specifically, Nakajima and Allen [4] examined the relationship between fundamental frequency (F0) and discourse structure in spontaneous task-oriented dialogue and found that F0 values tend to signal topic shift and

0-7803-2431-5/95 $4.00 © 1995 IEEE

**Table 1: Summary of Phrase-initial Usage of "Right" in VNS Corpus**

| "Right" used as: | Example | Phrase-initial occurrences | | |
|---|---|---|---|---|
| | | Used by Expert | Used by Traveller | Total |
| Direction | E: "Right on 27th" | 10 | 6 | 16 |
| Answer | T: "Then I turn left on Caremont?"<br>E: "Right." | 23 | 2 | 25 |
| Acknowledgment | E: "Turn left again heading north on Elizabeth."<br>T: "Right." | 4 | 10 | 14 |
| Other | T: "Right now I believe I'm on Platte." | 1 | 1 | 2 |
| Total | | 38 | 19 | 57 |

topic continuation across pause boundaries. Pitch accents mark salient material [3], which may be useful not only in interpreting the intention behind the utterance but also in locating critical content words for recognition purposes. Prosodic cues have been used successfully in disambiguating syntactic structure in professionally read speech ([5], [6]), but we still lack computational models for understanding prosodic cues in spontaneous speech.

In this study we tested the feasibility of using prosodic cues to aid in semantic disambiguation. In particular, we looked for prosodic cues that could be extracted and analyzed quickly and that would be robust in the context of a fallible pitch tracker operating in a fallible system. In the next section we describe our experiment. In section 3 we report our results, and in section 4 we discuss our findings and their potential application to spoken language understanding systems.

## 2. EXPERIMENT

The data for this study were drawn from the Vehicle Navigation System (VNS) corpus, a collection of task-oriented human-human dialogues taking place over cellular telephone. In these conversations, one conversant (the traveller) is attempting to drive to several different addresses and is consulting the other conversant (the expert) for directions. There are a total of two experts and 21 travellers navigating to three destinations each, with each destination requiring from one to three conversations between expert and traveller. The corpus comprises 93 brief conversations of 9 to 62 turns each. All conversants were native speakers of American English.

In this corpus the word "right" appears frequently and is used in several distinct senses. The most common of these were:

(a) "Right" as a direction, e.g., "Right on 27th."

(b) "Right" as an affirmative answer to an explicit question, e.g., "Then I turn left on Claremont?" "Right."

(c) "Right" as the acknowledgment of a contribution made by the other conversant, e.g., "Turn left again heading north on Elizabeth." "Right."

"Right" as an answer and "right" as an acknowledgment are intuitively similar, although they are technically different speech acts. As an answer, "right" conveys agreement; this usage is roughly equivalent to "yes." As an acknowledgment, "right" functions as a confirmation that the preceding utterance was understood but does not necessarily convey agreement. The use of acknowledgments assure conversants that the dialogue is on track and as such are important for a spoken language understanding system to recognize and to respond to appropriately [7].

We addressed the question of whether there were consistent prosodic differences in the way the word "right" was pronounced in each usage. To eliminate co-articulation effects, we included only occurrences that were turn or phrase-initial. That is, we considered only occurrences in which the word "right" was preceded by a pause or by the non-overlapping speech of the other conversant. Utterance-initial cases are also of particular interest from the standpoint of potential usefulness; at the point at which a system would be processing these occurrences, it would have relatively few syntactic or pragmatic cues from preceding words.

With these restrictions, we found a total of 57 occurrences in the VNS corpus. Two coders independently classified these according to the categorization described above. One coder worked from both audio tapes and a

**Table 2: Comparison of Difference in Pitch Changes by Category**

| Comparison | | p-value | Significance |
|---|---|---|---|
| "Right" used as: | Compared to: | | |
| Acknowledgments | Directions | 0.1007 | Trend,<br>not significant |
| Answers | Directions | 0.1113 | Trend,<br>not significant |
| Acknowledgments plus answers | Directions | 0.0375 | Significant,<br>confidence 0.95 |
| Acknowledgments | Answers | 0.4315 | Cannot<br>distinguish |

word-level transcription of the corpus. The other coder worked only from the transcription. The coders disagreed in only one case, for an agreement of 98%. This difference was resolved by discussion. Table 1 summarizes the usage distribution.

In devising a measure of intonation, we wanted a metric that would be usable in the context of a spoken language understanding system. In particular, it should be robust in the presence of pitch tracker inaccuracies caused by, for example, glottalization. After some preliminary tests, we settled on a simplified prosodic representation in which we divided the word into equal thirds and measured the change in average pitch between the first and last thirds of the word. This is an inexact measure of prosodic tune; it does not, for example, capture the variations described by Grigoriu et al [8]. However, our measure does capture a high-level notion of intonation as "rising," "falling," or "level," and it is easy and fast to compute from data readily available in a typical spoken language understanding system.

Pitch determination by a pitch tracker is often inexact. Glottalization or overlapping speech can cause the pitch tracker to return erroneous values. Nevertheless, these values are the ones that would be available in a working system, so we used actual pitch tracker results instead of trying to determine the "true" pitch values. We eliminated pitch values higher than 350 Hz and lower than 50 Hz, but made no further effort to correct pitch-tracker errors. When one of the values was eliminated, we substituted the average pitch of the middle third of the word. In one instance, no valid pitch value could be determined. This case was classified as having "level" intonation. In all other cases at least two pitch values were available, which leads us expect that this measure will be reliably available in a working system.

**Table 3: Discrimination Based on Pitch Change Alone**

| True Category | Numbe r Correct | Number Incorrec t | % Correct |
|---|---|---|---|
| Acknowledgments plus answers | 27 | 12 | 69% |
| Directions | 11 | 5 | 69% |
| Other | 0 | 2 | 0% |
| Total | 38 | 19 | 67% |

In this study, then, we compared the intonational characteristics of different usages of the word "right" using a simplified measure of intonation in which we compare the difference in pitch between the first and last thirds of the word. We hypothesized that we would find detectable differences in pitch patterns across the three usages.

## 3. RESULTS

We compared the pitch changes found in the three categories of phrase-initial or utterance-initial "right" using the Welch Modified Two-Sample t-test. We found significant differences in pitch patterns; the results are summarized in Table 2. When "right" was used as an acknowledgment or answer, it was more likely to be pronounced with a falling intonation. When used as a direction, "right" was more likely to occur with a rising intonation. The individual comparison of the acknowledgments and answers categories with the directions category showed only a suggestive trend. When the acknowledgments were grouped with answers, however, the combined categories showed a significant difference compared to the directions category. Acknowledgments could not be reliably distinguished from answers based on pitch change.

Although there is a statistically significant difference between the intonation of acknowledgments/answers and directions, the differences are not reliable enough to allow systems to use prosody alone to distinguish between these usages. In Table 3 we show the results of using pitch alone as a word sense discriminator. We chose flat intonation as the discriminator, assigning occurrences with level or falling intonation to the acknowledgment/answer category and rising intonation to the directions category. When this is done, we assign instances to their correct category 67% of the time.

## 4. DISCUSSION

The pitch-change differences seen in this study reflect the intuitive observation that "right" used as an acknowledgment or answer is likely to be uttered with a falling pitch, which is typical of American English sentence-final intonation [3]. In fact, 18 of the 39 acknowledgment/answer "rights" occurred as single-word turns. When "right" is used as a direction, it is usually the first word in a longer phrase and as such is more likely to exhibit a rising intonation.

How can this result help us build more robust systems? If pitch change alone is not an adequate discriminator, and if the prosodic differences merely reflect the tendency for acknowledgments and answers to be uttered as single-word utterances of phrases, then it could be argued that a syntactic analysis would serve the same purpose. That may be true — if the recognizer is able to accurately return the words used in the utterance. When the recognizer returns an incorrect word string, however, syntactic analysis could be misleading or ambiguous. In such cases, pitch cues could help choose between ambiguous interpretations or could help signal the potential error.

Taken as one of many potential cues such as the context of the preceding utterance or the syntactic analysis of the utterance, a simple measure of pitch change could be helpful. Where potentially ambiguous words such as "right" occur at the beginning of a longer utterance, for example, the direction of pitch change could serve as a confirming cue when analyzing ambiguous or erroneous recognizer output. Pitch cues could be used dynamically to guide an integrated system in which acoustic analysis is interleaved with a probabilistic language model, as described by Goddeau [9].

In future work we will continue exploring the interrelationships between various cues available to a spoken language understanding system. We plan to focus next on cues contributing to the recognition of acknowledgments or confirmations in spontaneous speech.

## ACKNOWLEDGEMENTS

The authors thank U S West for the use of the VNS corpus and Stephen Sutton for his analysis of the VNS corpus. This research was supported by the National Science Foundation.

## REFERENCES

[1] R. Cole, L. Hirschman, et al., "Workshop on Spoken Language Understanding," Oregon Graduate Institute Technical Report No. CS/E 92-014, 1992.

[2] S. McRoy, "Using Multiple Knowledge Sources for Word Sense Discrimination," *Computational Linguistics*, Vol. 18, No. 1, pp. 1-30, 1992.

[3] J. Pierrehumbert & J. Hirschberg, "The Meaning of Intonational Contours in the Interpretation of Discourse," in *Intentions in Communication*, P. Cohen, J. Morgan, & M. Pollack (Eds.), Chapter 14, pp. 271-311, Cambridge, MA:MIT Press, 1990.

[4] S. Nakajima & J. Allen, "A Study on Prosody and Discourse Structure in Cooperative Dialogues," Rochester Tech. Report No. TRAINS-TN93-2, Sept., 1993.

[5] A. Hunt, "A Generalised Model for Utilising Prosodic Information in Continuous Speech Recognition," *ICASSP '94*, pp. II-181-II-184, 1994.

[6] P. Price, M. Ostendorf, S. Shattuck-Hufnagel, & C. Fong, "The Use of Prosody in Syntactic Disambiguation," in *Proceedings of the Fourth DARPA Workshop on Speech and Natural Language*, ed. P. Price, 1992.

[7] D. G. Novick & S. Sutton, "An Empirical Model of Acknowledgment for Spoken-Language Systems," in *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics (ACL 94)*, pp. 96-101, 1994.

[8] A. Grigoriu, J. P. Vonwiller, & R. W. King, "An Automatic Intonation Tone Contour Labelling and Classification Algorithm," *ICASSP '94*, pp. II-181-II-184, 1994.

[9] D. Goddeau, "Using Probabilistic Shift-Reduce Parsing in Speech Recognition Systems," *Proceedings of the 1992 International Conference on Spoken Language processing (ICSLP 92)*, pp. 321-324, 1992.