ACOUSTIC-PROSODIC DISAMBIGUATION OF DIRECT AND INDIRECT SPEECH ACTS

Jill Suzanne Nickerson* and Jennifer Chu-Carroll†
*Harvard University, Cambridge, MA 02138, USA
†Lucent Technologies Bell Laboratories, Murray Hill, NJ 07974, USA

ABSTRACT

The same surface form used for a speech act meant to be interpreted literally may be used for a speech act that takes on additional indirect meaning. We investigated acoustic and prosodic differences between the realizations of such direct versus indirect speech acts. We conducted a production experiment with seven native General American English speakers, each engaging in fourteen pairs of dialogues designed to elicit realizations of direct and indirect readings of fourteen surface yes/no questions. Our analyses of the acoustic-prosodic features of these utterances show that 1.) utterances realized with a low boundary tone are substantially more likely to have an indirect reading, 2.) the main pitch difference between direct and indirect speech acts is realized in the f0 of the final high peak, and 3.) impoliteness in dialogue context contributes to greater final f0 rise in a particular subclass of utterances.

1. INTRODUCTION

Since the surface form of direct and indirect speech acts may be identical, other features must be used to disambiguate them so that a speaker's intentions can be interpreted correctly. For example, "*Can you move the couch?*" can be interpreted directly to question the hearer's ability to move the couch. Alternatively, it can represent an indirect speech act requesting the hearer to move the couch. Allen and Perrault [1] developed a plan-based model for deriving either the direct or indirect interpretation of a speech act by taking into account the hearer's beliefs of the speaker's possible goals and beliefs. Though, in many cases, contextual information is sufficient for inferring such beliefs and, thus, contributes to recognizing a unique interpretation of an utterance, there are cases in which ambiguity remains, even when the utterance is interpreted in context. Given previous research relating prosodic features to discourse interpretation (see e.g. [4, 6]), we investigate whether speakers attempt to convey speech act interpretations of utterances in their acoustic-prosodic realizations. To determine how this ambiguity is resolved in production, we examine acoustic-prosodic patterns based on three features: 1.) tone pattern of the nucleus, 2.) maximum f0 at the final high boundary tone, and 3.) average f0 of the stressed vowel carrying the nuclear pitch accent.

Our analyses show that, despite the fact that not all speakers use the same strategies to distinguish between direct and indirect speech acts, general inter-speaker trends do exist. Indirect speech acts are substantially more likely than direct ones to be realized with a low boundary tone. When utterances are realized with a nuclear L*HH% tone pattern, direct and indirect speech acts are distinguished by pitch differences in the f0 of the high boundary tone as opposed to the f0 of the stressed vowel carrying the nuclear pitch accent. Finally, impoliteness in dialogue context results in higher final f0 rise in a particular subclass of indirect speech acts.

The rest of this paper is organized as follows. In Section 2, after briefly defining basic terminology, we discuss the categories of speech acts used in this study. Section 3 explains the design of our experiment, which included designing dialogues to trigger direct and indirect interpretations of speech acts and performing a production experiment to elicit these speech act productions. In Section 4, we discuss our analyses of the production data as well as the conclusions we drew from this study.

2. DIRECT VS. INDIRECT SPEECH ACTS

A direct speech act is a speech act that is meant to be interpreted literally and has a single illocutionary force. For instance, "*Can you ski?*" uttered for the sole purpose of obtaining a yes/no-response, is a direct speech act. It has the literal meaning "I ask you whether you know how to ski." An indirect speech act, on the other hand, is a speech act that takes on meaning in addition to the literal one; it has more than one illocutionary force [8]. "*Can you help me?*", which is often uttered as a request for assistance, for example, is an indirect speech act. In addition to the literal meaning "I ask you if you have the ability to help me", it has the indirect meaning "I request that you help me" [2].

Searle [8] grouped some of the sentences that are typically used to convey indirect requests and other types of directives into six categories. Of these six categories, three contain question-types whose syntactic forms are also conventionally used as direct speech acts, provided appropriate contextual cues exist. These categories are of particular importance in this study. The first category contains questions that concern the hearer's ability to perform an action (CAN) such as "*Can you move the couch?*" The second category contains those questions which concern the hearer's doing an action (WOULD) such as "*Would you take him to dinner?*" The final relevant category contains questions such as "*Would you be willing to work for me?*" which concern the hearer's willingness to do an action (WOULD WILLING).

The experiments described in the rest of this paper identify acoustic-prosodic differences between realizations of direct and indirect speech acts, as well as among direct and indirect realizations of utterances in the three categories (CAN, WOULD, WOULD WILLING).

3. EXPERIMENTAL DESIGN

3.1. Dialogue Designs

The goal of our experiment was to elicit production data that would allow us to analyze the differences in acoustic-prosodic features of direct and indirect speech acts. We designed a total of fourteen surface yes/no questions, each of which fell into one of the aforementioned three categories: CAN, WOULD, and WOULD WILLING. Two dialogue contexts were constructed for each question: one that triggered a direct interpretation of the speech act and one with contextual cues to influence an indirect reading. The contexts were designed so that, for each pair of dialogues, features known to affect acoustic-prosodic realization of utterances, such as givenness and newness [3] and contrast [7] are identical in both dialogues.

Consider the following two dialogue contexts for the utterance "*Can you move it?*":

(1) A: What are you doing?
B: I'm trying to move the couch.
A: How do you think you're going to move it with that chair in the way?
B: If you can see that I'm having trouble, why don't you help me?
You're just standing there doing nothing.
Before I fall over the chair, **can you move it?**

(2) C: I've never seen that game before. It looks pretty fun.
D: It's a new game called "Tiles".
C: What's the red tile for? **Can you move it?**
D: No, you only move the blue ones. You can play the next game, I'll teach you.

In the context of dialogue (1), the highlighted key utterance should be interpreted as A requesting B to move the chair. In other words, the CAN question should be interpreted as an indirect speech act. On the other hand, contextual information in dialogue (2) suggests that the key utterance be interpreted literally as a yes/no question, i.e., as a direct speech act.

In addition to distinguishing between utterance classes for direct and indirect speech acts, the indirect speech acts were also controlled to fall into one of three modes: polite, neutral, and impolite. The mode of a dialogue was signaled by the dialogue context. For instance, in contrast to dialogue (1) which has an impolite mode, dialogue (3) presents the context for an indirect interpretation of a similar CAN question in a neutral mode.

(3) E: Now that I'm working full time, you're going to have to take on some more responsibilities.
F: Oh mom, I already do a lot around here.
E: A few more responsibilities won't hurt. I'm doing a lot of things right now.
F: What can I do to help you?
E: Well, **can you make your lunch?**

Table 1 displays the number of speech acts of each question-type and mode that were used in our production experiment.

Q-type	Direct	Indirect		
		Polite	Neutral	Impolite
Can	6	0	3	3
Would	3	1	2	0
Would Willing	5	2	3	0

Table 1: Distribution of speech act types and modes.

3.2. Data Collection

We conducted a production experiment with seven speakers (A-G), all of whom were native General American English speakers. In order to obtain production data consisting of the twenty-eight key utterances (a direct and an indirect reading of each of the fourteen utterances), we engaged each of the speakers in the twenty-eight prepared dialogues. The speaker played the role of the individual who uttered the speech act of interest, and a pre-recorded voice played the role of the other participant.

During the production experiment, the speaker sat in front of a display terminal in a sound-proof booth and was prompted with text in the context of the given dialogue. The speaker was instructed to speak as s/he naturally would. The dialogues that the speaker saw contained no final punctuation (i.e., the surface yes/no questions of interest were not followed by question marks), and one of the roles was in all capital letters. The speaker was asked to play the role of the participant whose utterances were in all capital letters, and a pre-recorded voice, which was broadcast through a speaker system in the booth at the appropriate times, played the role of the other participant. Only the voice of the speaker was recorded. The speaker had the option of re-recording. Also, the dialogues were pseudo-randomized, so the speaker would not participate in a dialogue with the same key utterance as the dialogue immediately preceding it. This process was repeated for each of the seven speakers, thus resulting in ninety-eight pairs of direct and indirect speech acts.

4. DATA PROCESSING AND ANALYSIS

After collecting the production data, each speaker's direct and indirect readings of the fourteen key utterances were isolated. The f0 contour, waveform, and spectrogram of these key utterances were computed and examined using the Entropics speech analysis software. We then analyzed the acoustic-prosodic features of these key utterances.

4.1. Speech Act Disambiguation Based on Prosodic Analysis

For each speaker, we analyzed the intonation contour of each speech act production. Table 2 shows the distribution of tone patterns for each speaker. The majority of the speech act productions exhibited the tone pattern typical of yes/no questions: L*HH%, one with terminal rise [3].[1] However, utterances realized with a low boundary tone were substantially more likely to be intended as indirect than direct speech acts: 73% of the utterances with low boundary tones (represented in the column 'Other') had indirect readings.

Speaker	Tone Patterns of Speech Act Productions	
	L*HH%(Indirect/Direct)	Other (Indirect/Direct)
A	14 / 14	0 / 0
B	14 / 14	0 / 0
C	11 / 13	3 / 1
D	9 / 12	5 / 2
E	9 / 11	5 / 3
F	11 / 13	3 / 1
G	11 / 14	3 / 0

Table 2: Distribution of tone patterns.

4.2. Speech Act Disambiguation Based on Acoustic Analysis

The distribution of tone patterns in Table 2 shows that, of all surface yes/no question realized with an L*HH% tone pattern, 46% had indirect readings while 54% had direct readings. To further disambiguate this class of utterances, we investigated some acoustic features of these utterances. For each such utterance realized with the L*HH% tone pattern, we labeled three points on the pitch track: location of the high boundary tone, as well as the beginning and end of the stressed vowel of the nuclear pitch accent, using information from the spectrogram.

We analyzed the acoustic properties of these utterances along two dimensions. First, we attempted to find speaker-independent distinctions between direct and indirect speech acts among the three classes of utterances (CAN, WOULD, and WOULD WILLING), as well as within the three modes (polite, neutral, and impolite). Next, we analyzed individual speakers to discover speaker-specific strategies for speech act realization.

4.2.1 Inter-speaker Analysis. We first analyzed the pitch differences between the direct and indirect speech acts across all speakers. For this analysis, we selected pairs of utterances that satisfied the following criteria: 1.) both utterances were realized with the L*HH% tone pattern, and 2.) both utterances had the same stressed vowel. We compared the f0 of the boundary tone (henceforth referred to as %f0) and the average f0 of the vowel carrying the nuclear pitch accent (henceforth referred to as *f0). Figure 1 shows that for most speakers, the *f0's are roughly equivalent for both direct and indirect readings of the same utterance, while the %f0's differ quite substantially for the two different readings. For each pair of direct and indirect readings of the same sentence, we computed the absolute differences of the %f0 values (BOUNDARY_DIFF) as well as of the *f0 values (STRESS_DIFF). T-tests performed on BOUNDARY_DIFF and STRESS_DIFF show that %f0 differences are significantly greater than *f0 differences ($p < .001$), indicating that pitch differences between direct and indirect speech acts are realized in the f0 of the boundary tone as opposed to the f0 of the stressed vowel.[2]

Next, we analyzed whether there were significant differences between the ways speakers distinguished between the direct and indirect readings of utterances based on their utterance classes. For this analysis, we used all pairs of utterances where both readings were rendered with the L*HH% tone pattern. Figure 2 shows a scatter plot of the %f0 of each pair of utterances, distinguished by the utterance class. We computed the difference in %f0 between the direct and indirect readings of each pair of utterances, and the differences among the three classes are shown to be marginally significant (ANOVA test, $p < .06$).

We further analyzed whether the distinction between modes (polite, neutral, and impolite) within each utterance class contributed to different acoustic realization for direct and indirect speech acts. For each utterance type, we grouped all pairs of utterances based on the mode of the dialogue context of the indirect reading (as shown in Table 1). Within each group, we computed the difference in %f0 for each pair of direct and indirect speech acts, and performed a t-test to determine, for each utterance type, whether the f0 differences for the groups are significantly different. Table 3 shows that for the CAN utterances, the neutral and impolite dialogue modes resulted in significant differences in the realization of the %f0 between direct and indirect interpretations of these utterances. On the other hand, for the other two utterance classes, the dialogue mode did not have such an effect.

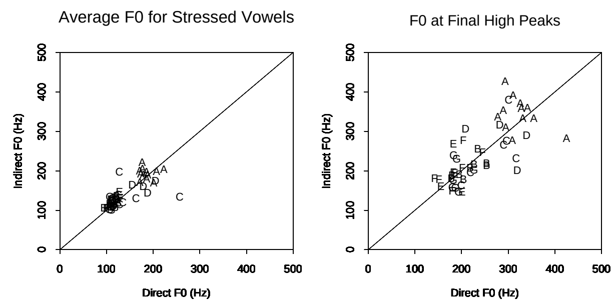

Figure 1: Comparison between average f0 of stressed vowels (*f0's, left panel) and f0 at high boundary tones (%f0's, right panel) for direct vs. indirect speech acts across speakers (A-G).

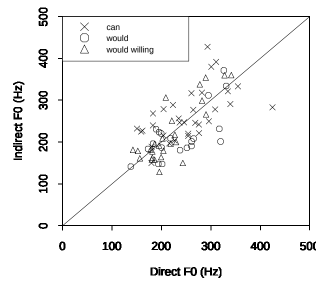

Figure 2: Comparison of %f0 for direct vs. indirect speech acts based on utterance class.

Q-type	df	Relationship between %f0	p
Can	12	impolite > neutral	**<.001**
Would	12	polite > neutral	< .24
Would Willing	12	polite > neutral	<.42

Table 3: T-test results for different dialogue modes for each utterance class.

4.2.2. Intra-speaker Analysis. The results presented in the previous section were obtained by considering all speakers as a group. We were further interested in examining whether individual speakers adopt different strategies to distinguish between direct and indirect speech acts. In particular, in the previous section, we have shown that the pitch differences between direct and indirect speech acts are realized in the f0 of the boundary tone, as opposed to in that of the stressed vowel when considering all speakers as a group. In this section, we first examine whether or not the same result can be reproduced when considering each speaker individually. Next, we examine in further detail the patterns each speaker adopts in distinguishing between direct and indirect speech acts.

To examine whether individual speakers exhibited the trends discovered in our inter-speaker analysis, we repeated the first analysis in Section 4.2.1 in which we performed a t-test on the absolute differences between indirect and direct %f0's and the absolute differences between indirect and direct *f0's. The results show that for five out of seven speakers, the same conclusion drawn from the earlier inter-speaker analysis applies at $p < .05$. For one speaker, the same conclusion can be drawn at $p < .09$, while for the last speaker, the observation does not apply.

One interesting observation in Figure 1 is that the points representing %f0 for indirect speech acts vs. %f0 for direct speech acts for the same speaker occurred on either side of the diagonal line. This prompted us to investigate, in addition to realizing indirect speech acts differently from their direct counterparts at the final high rise, whether one reading was uniformly produced with a higher boundary f0 than the other. For this analysis, we included all utterances realized with the L*HH% tone pattern. For each speaker, we first performed a simple analysis of the utterance pairs by categorizing them into two groups, depending on whether the indirect %f0 was realized higher than the direct %f0 or vice versa. The results of this analysis, summarized in part of Table 4, show that some speakers (such as A and F) fairly uniformly produced one interpretation with a higher %f0 than the other, while other speakers (such as G) made no such distinction. Next, we examined the statistical significance of individual speaker strategies by computing the difference between the %f0 for the indirect reading and that for the direct reading for each utterance pair and performed a one-sample t-test on the computed differences. The results of this analysis, shown in the rest of Table 4, indicate that three out of the seven speakers adopted marginally uniform strategies in distinguishing between direct and indirect speech acts (speakers A and F realized indirect speech acts with higher %f0 while speaker C realized them with lower %f0), while the other speakers basically made no measurable distinction.

Speaker	High %f0 count (Indirect/Direct)	T-test results		
		df	Rel between %f0	p
A	10 / 4	13	indirect > direct	**<.15**
B	5 / 9	13	direct > indirect	<.31
C	3 / 8	10	direct > indirect	**<.08**
D	3 / 6	8	direct > indirect	<.30
E	5 / 3	7	direct > indirect	<.40
F	9 / 2	10	Indirect > direct	**<.10**
G	5 / 6	10	Indirect > direct	<.49

Table 4: Results for individual speaker analysis.

5. CONCLUSION

We examined acoustic-prosodic features of direct and indirect speech act pairs with the same surface form by conducting a production experiment. We showed that the two types of speech acts have different acoustic-prosodic features. Utterances realized with a low boundary tone are more likely to have indirect readings. Speakers differentiated direct and indirect speech acts with terminal rise by pitch differences in the f0 of the boundary tone as opposed to the average f0 of the stressed vowel carrying the nuclear pitch accent. "Impolite" CAN questions were distinguished from "neutral" CAN questions by a higher final f0 rise. In addition, the three question types used in the production experiment differed marginally from each other by their final f0 rises. Though these overall trends were observed across all speakers, consistency within individual speakers to the same effect is less apparent.

There is still a lot to be learned concerning which acoustic-prosodic features speakers and hearers use and how they use these features to distinguish between direct and indirect speech acts. Though acoustic-prosodic factors alone may not be able to fully disambiguate whether a speech act production is meant to be interpreted as a direct or an indirect speech act, the general trends that we observed across speakers in the production data that we collected warrant further studies. An accurate characterization of the disambiguating features that speakers use to produce a speech act with the desired intention(s) and that hearers use to correctly interpret these intention(s) would pave the way for improvements in systems for interpreting and generating dialogues and in applications such as concept-to-speech synthesis.

ACKNOWLEDGMENTS

We wish to thank Jennifer Venditti for her valuable suggestions and instructive comments. We thank Julia Hirschberg for suggesting this topic for investigation, Jan van Santen for helpful discussions on statistical analysis, Chilin Shih and Michael Tanenblatt for their help with speech recording, and Bernd Moebius for general discussions on acoustic phonetics. We are grateful to Bob Carpenter, Mark Core, Jim Hieronymus, Gerald Penn, Christer Samuelsson, and Chilin Shih for helpful discussions, as well as to Jennifer Venditti, Christine Nakatani, and Bob Carpenter for their comments on an earlier draft of this paper. Last but not least, we would like to thank our subjects for taking the time to participate in our production experiment.

NOTES

1. In the production data we collected, all key utterances were realized by one intonational phrase. Thus, the nuclear accent and boundary tone combinations we examined occur at the end of the speaker's utterances in all cases.
2. This result is consistent with previous claims that peak variations are greater than low variations [5]. In addition to supporting this claim, our results suggest that the peak variations could convey a difference in utterance meaning.

REFERENCES

[1] Allen, James F. and Raymond C. Perrault 1980. Analyzing Intention in Utterances. In *Artificial Intelligence*, 15, 143-178.

[2] Clark, Herbert H. 1979. Responding to Indirect Speech Acts. In *Cognitive Psychology*, 11, 430-477. New York: Academic Press.

[3] Hirschberg, Julia 1990. Accent and discourse context: assigning pitch accent in synthetic speech. In *Proceedings of the Eighth National Conference on Artificial Intelligence*, 952-957.

[4] Liberman, Mark and Ivan. A. Sag 1974. Prosodic Form and Discourse Function. In *Papers from the Tenth Regional Meeting*, Chicago Linguistic Society, Chicago, IL.

[5] Liberman, Mark and Janet Pierrehumbert, 1984. Intonational Invariance under Changes in Pitch Range and Length. In M. Aronoff and R. Oehrle (eds.), *Language Sound Structure*, 157-233. MIT Press.

[6] Pierrehumbert, Janet and Julia Hirschberg 1990. The Meaning of Intonational Contours in the Interpretation of Discourse. In P.Cohen, J. Morgan, and M. Pollack (eds.), *Intentions in Communication*, 271-311.

[7] Prevost, Scott 1996. Modeling Contrast in the Generation and Synthesis of Spoken Language. In *Proceedings of the Fourth International Conference on Spoken Language Processing*, 1349-1352.

[8] Searle, John R. 1975. Indirect Speech Acts. In Peter Cole and Jerry L. Morgan (eds.), *Syntax and Semantics*, vol.3: *Speech Acts*, 59-82.