Interpreting Final Rises: Task and Role Factors

Catherine Lai

Centre for Speech Technology Research,
School of Informatics, University of Edinburgh, United Kingdom
clai@inf.ed.ac.uk

Abstract

This paper examines the distribution of utterance final pitch rises in dialogues with different task structures. More specifically, we examine map-task and topical conversation dialogues of Southern Standard British English speakers in the IViE corpus. Overall, we find that the map-task dialogues contain more rising features, where these mainly arise from instructions and affirmatives. While rise features were somewhat predictive of turn-changes, these effects were swamped by task and role effects. Final rises were not predictive of affirmative responses. These findings indicate that while rises can be interpreted as indicating some sort of contingency, it is with respect to the higher level discourse structure rather than the specific utterance bearing the rise. We explore the relationship between rises and the need for co-ordination in dialogue, and hypothesize that the more speakers have to co-ordinate in a dialogue, the more rising features we will see on non-question utterances. In general, these sorts of contextual conditions need to be taken into account when we collect and analyze intonational data, and when we link them to speaker states such as uncertainty or submissiveness.

Index Terms: Intonation, task-oriented dialogue, rises.

1. Introduction

The question of what prosody contributes to meaning is a key problem for both automated spoken language understanding and theories of semantics and pragmatics. In particular, a good number of studies have investigated how utterance final pitch rises and falls relate to epistemic and affectual states of speakers and how this relates to tasks such as dialogue move detection. Such studies generally examine how prosody affects the interpretation of the carrier utterance in the immediate context, e.g. whether a cue word is interpreted as a backchannel or not. While the local context clearly has a large effect on how prosody is interpreted, we would also like to know what impact higher level features such as task and role have as well.

One reason we expect higher level features to affect the interpretation of prosody follows from the incongruence of findings based on single dialogue types. For example, a correlation between rises and backchannels has been reported in map task dialogues in Bari Italian [1], Swedish [2], and Dutch [3], as well as in other game corpora in English [4, 5, 6]. However, these sorts of results are absent from studies of more free-form conversational dialogues in English [7, 8, 9] and Hindi [10]. We would like to know whether the differences in rise distributions from these backchannel studies extend to other sorts of dialogue moves, and if so, why.

To examine this, we look at the IViE (Intonational Variation in English) corpus [11]. The corpus contains speech of various styles including isolated read sentences as well as spontaneous conversational and task-oriented dialogues (the map task) from speakers of urban regions of the United Kingdom. In this paper we look at the boundary intonation of speakers from Cambridge (i.e. Standard Southern British English) in these different modes of speech. The motivation for looking at this dialect in particular is that out of those included in the corpus the intonation pattern for this region's declarative statements has been found to be the most pervasively falling or low at the boundary in read speech. Thus, rises are more likely to be seen as deviations from the canonical. When we observe rises, we expect them to mean something more than just a phonological boundary. So, looking at this data enables us to look at the effects of task and role on the frequency of rises, as well as giving as a more general view on how task-oriented and conversational dialogue differ.

2. Background

Direct comparisons of conversational and task-oriented speech have mainly focused on the greater need for affectual/emotional modelling in the former [12, 13]. While automatic role recognition has received more attention recently [14, 15], studies have not generally investigated in any detail how prosody varies with different role/move categories. However, some investigations of this type have been carried out with the goal of improving expressive speech synthesis. For example, [16] find that 'Assess' moves in the AMI corpus were produced with tenser voice quality, while project managers had higher average F0 and vocal effort. Dominant participants exhibited 'louder' voice quality features in [17]. While these studies provide broad descriptions for specific roles, they don't look at the contribution of intonation features like terminal rises in any detail.

Theoretical studies have analyzed rises as expressing uncertainty [18, 19, 20] or submissiveness [21, 22]. So, we might expect to find more rises in the productions of participants in socially submissive roles rather than leader type roles. However, empirical studies suggests that the distribution of rises depends heavily on situational and cultural conventions. For example, in a qualitative analysis of sorority speech, rises were used by senior members to take and hold the floor in monologues, while they were perceived as expressing uncertainty in narratives by uninitiated members of the group [23]. Similarly, in a comparison of several dialogue types, rises were found to be more prevalent in dialogues where one person has a socially dominant role, e.g. academic supervision versus informal office conversations [24] . Moreover, it was the socially dominant participant who produced the rises.

The latter study doesn't conditionalize over different move types, so it's not clear whether the more one-sided conversations simply involve more questions. In particular we would like to know when rises occur on sentence types that are canon-

10.21437/SpeechProsody.2014-91

ically falling in the dialect we are examining, e.g. declarative statements (informing moves) and imperatives (instructions) [25, 11]. Rises have also been analyzed in terms of contingency [26, 27], hearer dependence [28] about the rise carrying utterance. So, we would like to know whether the distribution of rises in a dialogue can be explained in terms of whether moves need explicit ratification or not. This would predict a higher number of affirmatives following rising moves. More generally, we would like to know if the distribution of rises can be adduced from the turn-taking structure of the dialogues, i.e. whether rises give or hold the floor.

3. Experimental Setup

3.1. Data

The IViE corpus was developed to systematically study differences across regions, speakers, and styles [11]. As mentioned previously, we look at data from Cambridge speakers as the most consistently 'falling' dialect in the collection. Twelve speakers (6 male, 6 female) from each region were recorded between 1997-2000. The speakers were 16 years old at the time of recording and had been born in and grown up in the region. The recordings include a mixture of read and spontaneous speech, of which we use the following:

- *Map task (map):* Each participant was given a map of a small town. Participants took one of two roles: Instruction *giver* and *follower*. The goal was for the giver to explain a pre-defined route around town on their map to the follower, who traced it out on their own map. The task ended when the route was completed to the satisfaction of both participants. Maps differed in place names and locations of landmarks, so speakers had to work to establish common ground. Speakers were separated by screens so they could not see each other. More details about this task can be found in [29].
- *Free conversation (conv):* Participants discussed smoking, face-to-face. Speakers had the same role, which was simply a *participant*

3.2. Segmentation and Annotations

Only short portions from four speakers were transcribed and annotated at sentence type level for each of these dialogue types in the official IViE release, so additional annotation was undertaken. The dialogues were manually segmented into utterances corresponding to whole meaning units rather than phonological phrases (cf. [30]). This was done as a conservative measure of the frequency of rises. Sentential segmentation delimited whole propositions including any embedding. Similarly, imperative utterances mapped to one action (i.e. one segment of the route). A number of sub-sentential clauses also formed separate utterances, e.g. an NP or VP as an answer or a modifier separated by an affirmative, which were as marked as XPs. Utterances were labelled with sentence (syntactic) and move types:

- Sentence type: Declarative (dec), Imperative (imp), Polar question (yno), Wh-question (whq), Tag (tag) question, Affirmative (affirm), Negative (neg), Cue word (cw), If antecedent (IFA), XP (XP).
- Broad dialogue moves: Affirm, Neg, Contra (direct contradictions), CW (cue words), Inform, Instruct, Q (non-syntactically marked question), YNQ (polar question), WhQ (wh-question), Tag (Tag-question), sync (synchronize).

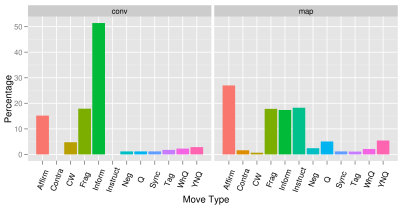

Figure 1: *Proportions of moves:* $N_{conv} = 430$, $N_{map} = 1287$.

The rationale for using such broad move types was to keep the annotation in terms of easily identifiable categories which could be refined in the future. In many cases, one sentence type dominated a move type (e.g. wh-questions). The main points of variation were in the declaratives which we see as instructions, informing moves, and questions, amongst other moves. Similarly, instructions, while primarily imperative in form, were also expressed as declaratives, polar questions or if-antecendents ('If you could go to the church'). The *sync* category captured utterances in the map task like 'You should be at the Anne's Arms' which were not quite questions, instructions or inform moves. A backchannel category was not included as it was not clearthat the distinction could be reliably made [31]. Moreover, it did not seem that any of the affirmatives in the map task could be clearly classified as simple signals of attention. In the investigations to follow we will concentrate on the most populous and easy to identify categories: Affirmative, Instructions and Inform moves. Utterance segmentation of the dialogues resulted in 430 and 1287 utterances for the conversational and map task sets respectively. The distribution of moves is shown in Figure 1.

3.3. Boundary pitch features

The target area for analysis was the stretch of speech from the last prominence rather than the last word. Extension away from the last word was generally due simply to stress assignment in compounds (e.g. *bowling* alley) or deaccenting of pronouns, (e.g. *about* it). Utterances with speaker overlap at the target were excluded from the prosodic analysis (3% conv, 4% map).

The F0 contour data on target segments was extracted using the Praat autocorrelation method. Parameter settings were automatically determined using the method described in [32]. Utterances which produced less than 5 F0 points were discarded. The F0 data was normalized into semitones relative to the median F0 value (Hz) for each speaker using data produced in all IViE tasks including read speech [11]. F0 contours were smoothed using a Butterworth filter and contours were approximated using Legendre polynomial decomposition of order 4 (cf. [33]). Instead of making categorical judgements about shape, we will instead look at first three coefficients where LC1 increases with overall pitch *height*, LC2 increases with positive contour slope (or *tilt*), and LC3 with *convexity*. Positive LC2 and LC3 indicate a fall-rise contour with an overall rising trend, negative LC2 and LC3 indicate a rise-fall contour, while values close to the origin indicate a flat contour. Previous work has described the relationship between these features and ToBI perceptual labels [34].

4. Results

4.1. Rises across dialogue types

Figure 2 shows the distribution of values for LC2 ad LC3 by move type. Overall, the map-task data can be characterized as

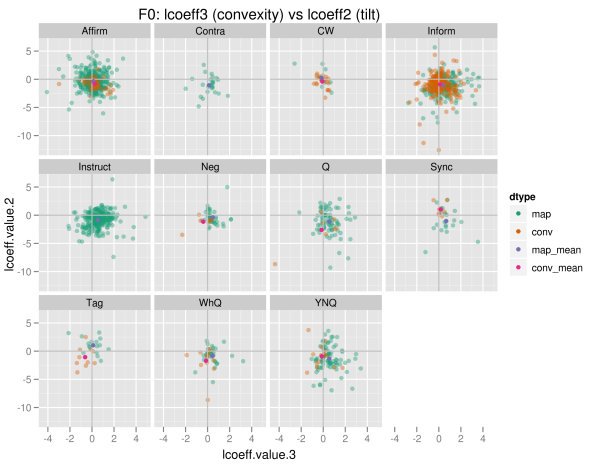

Figure 2: *Tilt (LC2) and Convexity (LC3), by move type, with means.*

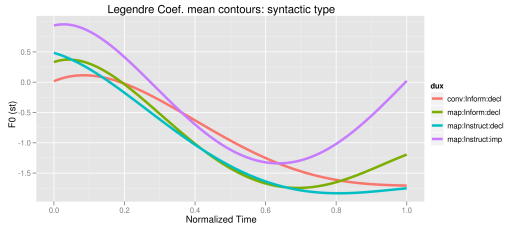

Figure 3: *Mean contours Instruct and Inform moves based on Legendre coefficients grouped by syntactic type.*

having more rising features, with these mainly coming from instructions and affirmatives. We see that Instruct moves are mostly situated in the positive LC3 space, indicating a fall-rise contour. The distribution of affirmatives in the map task extends into the positive LC2 space, indicating rising tilt.

Inform and Instruct moves make up 44% of the utterances in the map task, while 66% of conversation moves were Informs. These moves provide most of the 'new' content in the dialogue and are canonically falling in Southern Standard British English [25]. So, this subset of moves are good indicators that task affects the distribution of rising features. Figure 3 shows mean contours for Inform and Instruct moves which are declaratives, as well as the imperatives for comparison. Within the Instruct moves, syntactic imperatives are particularly rising compared to declarative instructions. We see that, on average, Inform declaratives are more rising in the map task than in conversational speech.

In order to quantify this we model the relationship between Legendre polynomial coefficients and dialogue factors (role, move, and syntactic type).[1] We fit (non-nested) multilevel linear models predicting values of LC1, LC2 and LC3. The model parameters were estimated using the package `lmer` in `R`. We only see significant effects for role when we look at convexity (LC3). The effect estimates and confidence intervals shown in Figure 4 indicates that being an instruction giver (1) increases convexity, while simply being a conversational participant (3) decreases it. In terms of moves, there is a signficant positive effect on convexity for imperatives (i.e. fall-rise). The effect of

[1]Note: Role encodes the task.

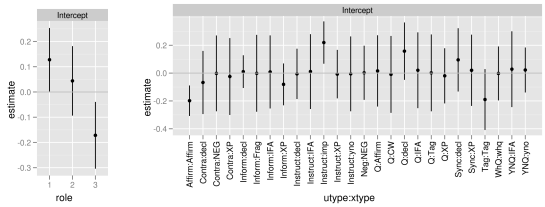

Figure 4: *F0 convexity (LC3) parameter estimates.*

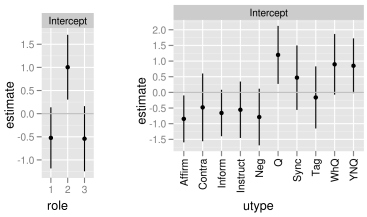

Figure 5: *Speaker change parameter estimates: 1=giver, 2=follower, 3=conversational participant.*

Affirm moves was negative on convexity, but positive on height and tilt, with a positive value for the instruction follower. This again points to there being more rising affirmatives in the task-oriented dialogue. Interestingly, yes/no and declarative questions have a negative relationship with tilt. This suggest that the specific questioning use of rises is less in play in these sorts of dialogues.

4.2. Turn-taking

Since Cambridge imperatives and declaratives are usually described falling at the boundary we would like to know whether the rises we see in the map task data can be attributed to other discourse factors like turn-taking. To do this, we fit parameters for multilevel logistic regression models (stay=0, switch=1). We compare a model containing speaker, role and move factors with one extended with Legendre coefficients as predictors (adding syntactic indicators did not improve the model fit).

Looking at the parameter estimates in Figure 5, we see that being the instruction follower (role 2) increases the probability of a switch by 25%. The trend for the other two roles is to hold the floor. In terms of move type, we see that the broad class of question moves increase the probability of switching, while content adding moves decrease it, although we saw previously that these generally had a falling tilt. With respect to move-role interaction, affirmatives produced by the instruction giver are likely to result in stays. That is, instruction givers seem to use affirmatives as a ready signal. The effects of the other move-role combinations are quite small in comparison.

Figure 6 shows significant positive effects for LC2 and LC3 (estimated coefficients: 0.12 and 0.2 respectively.) However, the magnitudes of these effects are relatively small compared to the effects of role and move. For example, when LC2 equals 1 we get an approximate 3% increase in probability of a switch, where the 95% interval of values in the observed data for LC2 is $(-4.15, 2.30)$ (LC3: $(-1.39, 2.12)$). So it seems that having higher tilt or convexity nudges up the probability of a speaker switch, but the contribution is not as strong as that of move or role. As we would expect, question type moves are generally turn giving irrespective of whether the utterance has a rising or falling boundary. To see whether rising features have different

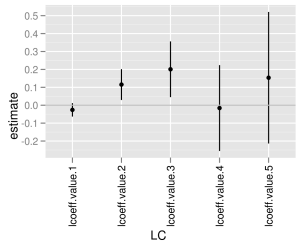

Figure 6: *Turn-taking: LC data as individual level predictors.*

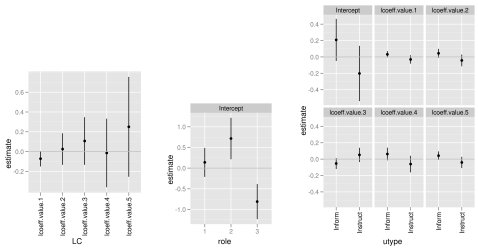

Figure 7: *Predicting Affirmatives after Inform/Instruct moves with LC features as individual level predictors and predictors on the move group.*

effects on different move types for turn-taking we extend the model to allow coefficients for the LC values to vary by move type. However, the difference between the previous two models is very small: DIC decreases from 1809 to 1802 where the new model adds many more parameters. So, it does not appear that rising features have much of a role to play in determining who takes the floor.

4.3. Eliciting Affirmatives

To check whether rising features signal a need for explicit ratification/agreement, we fit a multilevel logistic model predicting whether or not Inform or Instruct moves are followed by Affirm moves with the same predictors as above but coding an Affirm response as 1, and other responses as 0. If rises do signal a need for ratification we would expect to see that the probability of an affirmative increase with LC2 and LC3. After controlling for the higher level dialogue features, we see that the effects of the contour shape features to be, again, dwarfed by the effect of role. Estimated coefficients for the pitch features are around ± 0.05 at the move level, resulting in about a $\pm 1\%$ change in the likelihood of an affirmative response for every LC coefficient unit (cf. Figure 7). The effects are similarly close to zero at the individual level. On the other hand, being the map task follower, again, increases the probability of the next move being an affirmative by about 18%. In general, we don't see that pitch shape features on content moves are predictive of whether or not that move will be explicitly ratified.

5. Discussion

Our goal was to find out if higher level effects like task, role and move type had an effect of whether an utterance was produced with rising features. Looking at the Cambridge IViE data, it seems that we do get more rising features in the task-oriented speech than the free conversation, mostly with respect to instructions and affirmatives. While we saw that significant positive effects of tilt and convexity coefficients for speaker changes and production of affirmatives, we also saw that these effects were dwarfed by the effect of role. So, whatever the rising features are doing on these turns, it does not seem that their main role is to manage turn-taking. Instead, turn-taking strategy seems mostly dictated by the higher level, task structure of the dialogue.

How do our results fit with other analyses of rises? It is hard to reconcile the data with accounts that link rises to *propositional* attitudes (i.e. attitudes about the content bearing the rise). For example, we wouldn't want to associate rising features with how we usually think of epistemic uncertainty (contra [18]'s Maxim of Pitch), nor would be want to associate them with social submissiveness [21] or lack of speaker commitment [28], since these features are predominantly used by by the instruction giver, i.e. the situationally dominant speaker. In fact, if the instruction giver is uncertain about anything, it's not about the actual route. Instead it seems to be about whether the follower can or will carry out the task, i.e. discourse structural uncertainty. Similarly, the instruction giver is at the mercy of the follower in terms of task completion. So, at a glance we might say that the map-task shows more rising features because it just has more contingent elements (cf. [27, 26]. However, we saw that rises don't seem to elicit ratification in terms of explicit affirmative responses. So, if rises do signal contigency it is not necessarily about the current utterance.

An alternative is that the speaker deploys rising features because it is important to attend to whether a task is open or closed, since each subtask is dependent on the subtask before it. That is, we get more rises because there is a more well defined subtask structure (cf. [30]) and participants need a high level of common ground co-ordination in order to reach the end-goal of the dialogue. The need for co-ordination is much less pressing in conversational speech where participants are basically voicing opinions. So, the notion of contingency we are dealing with is at a higher level than accepting single instructions. From this point of view, explicit affirmation (rise or not) is a good strategy for the map-task follower, but rises primarily signal that there is more to come [35]. Thus we expect to see more rises in dialogues where greater quality of co-ordination is required.

6. Conclusion and Future Work

In this paper we saw that that higher level discourse factors, like task and role, have an effect on whether an utterance is produced with rising features or not. Overall, we found that content providing utterances in map-task dialogues had greater convexity than those from the conversational dialogue. Most of this seemed to come from instruction moves which often had a distinct fall-rise shape. While the rate of Affirmative moves was higher in the map task, we didn't find any strong link between rising features – higher LC2 and LC3 – and affirmative responses, or more generally speaker switches or stays. This state of affairs sits best with discourse structural analyses of rises, rather than notions like submissiveness or uncertainty. It appears that the more speakers have to co-ordinate through verbal signals, the more rising features we expect to see. So, these sorts of contextual conditions need to be taken into account when we collect and analyze intonational data.

Further work looking at the relationship between frequency of rises and the overall quality of task-completion, as well as comparision to other dialects, especially default rising ones such as Belfast English, and styles, will help complete the picture of where intonation fits into semantic/pragmatic theories.

7. References

[1] M. Savino, "The intonation of backchannels in italian task-oriented dialogues: cues to turn-taking dynamics, information status and speakers attitude," in *5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, 2011.

[2] M. Heldner, J. Edlund, K. Laskowski, and A. Pelcé, "Prosodic features in the vicinity of silences and overlaps," in *Proc. 10th Nordic Conference on Prosody*. Citeseer, 2008, pp. 95–105.

[3] J. Caspers, "Melodic characteristics of backchannels in dutch map task dialogues," in *Sixth International Conference on Spoken Language Processing*, 2000.

[4] B. Hockey, "Prosody and the role of okay and uh-huh in discourse," in *Proceedings of the eastern states conference on linguistics*. Citeseer, 1993, pp. 128–136.

[5] S. Benus, A. Gravano, and J. Hirschberg, "The prosody of backchannels in American English," in *Proceedings of ICPhS 2007*, 2007, pp. 1065–1068.

[6] A. Gravano, "Turn-Taking and Affirmative Cue Words in Task-Oriented Dialogue," Ph.D. dissertation, Columbia University, 2009.

[7] E. Shriberg, R. Bates, P. Taylor, A. Stolcke, D. Jurafsky, K. Ries, N. Coccaro, R. Martin, and M. Meteer, "Can Prosody Aid the Automatic Classification of Dialog Acts in Conversational Speech?" *Language and Speech*, vol. 41, no. 3-4, pp. 443–492, 1998.

[8] N. Ward, "Pragmatic functions of prosodic features in non-lexical utterances," in *Speech Prosody 2004*, vol. 4, 2004, pp. 325–328.

[9] K. Truong and D. Heylen, "Disambiguating the functions of conversational sounds with prosody: the case of'yeah'," in *Proceedings of Interspeech 2010*. International Speech Communication Association (ISCA), 2010.

[10] S. Prasad and K. Bali, "Prosody cues for classification of the discourse particle" hã" in hindi," in *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[11] E. Grabe, "Intonational variation in urban dialects of english spoken in the british isles," in *Regional variation in intonation*, P. Gilles and J. Peters, Eds. Linguistische Arbeiten, Tuebingen, Niemeyer, 2004, pp. 9–31.

[12] N. Campbell, "Developments in corpus-based speech synthesis: Approaching natural conversational speech," *IEICE transactions on information and systems*, vol. 88, no. 3, pp. 376–383, 2005.

[13] Y. Wilks, R. Catizone, S. Worgan, and M. Turunen, "Review: Some background on dialogue management and conversational speech for dialogue systems," *Computer Speech and Language*, vol. 25, no. 2, pp. 128–139, 2011.

[14] A. Vinciarelli, F. Valente, S. Yella, and A. Sapru, "Understanding social signals in multi-party conversations: Automatic recognition of socio-emotional roles in the ami meeting corpus," in *2011 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2011, pp. 374–379.

[15] S. Renals, T. Hain, and H. Bourlard, "Recognition and understanding of meetings the ami and amida projects," in *Automatic Speech Recognition & Understanding, 2007. ASRU. IEEE Workshop on*. IEEE, 2007, pp. 238–247.

[16] M. Charfuelan and M. Schröder, "Investigating the prosody and voice quality of social signals in scenario meetings," *Affective Computing and Intelligent Interaction*, pp. 46–56, 2011.

[17] M. Charfuelan, M. Schröder, and I. Steiner, "Prosody and voice quality of vocal social signals: the case of dominance in scenario meetings," in *Interspeech'10*, 2010.

[18] J. Hirschberg, "Communication and prosody: Functional aspects of prosody," *Speech Communication*, vol. 36, no. 1-2, pp. 31–43, 2002.

[19] M. Nilsenova, "Rises and falls. studies in the semantics and pragmatics of intonation," Ph.D. dissertation, University of Amsterdam, 2006.

[20] B. Reese, "Bias in questions," Ph.D. dissertation, University of Texas at Austin, 2007.

[21] A. Merin and C. Bartels, "Decision-Theoretic Semantics for Intonation," Universitt Stuttgart and Universitt Tübingen, Tech. Rep. Bericht nr. 88., 1997.

[22] C. Gussenhoven and A. Chen, "Universal and Language-Specific Effects in the Perception of Question Intonation," in *Sixth International Conference on Spoken Language Processing*. ISCA, 2000.

[23] C. McLemore, "The pragmatic interpretation of english intonation: Sorority speech," Ph.D. dissertation, University of Texas at Austin, 1991.

[24] W. Cheng and M. Warren, "//CAN i help you //: The use of rise and rise-fall tones in the Hong Kong Corpus of Spoken English," *International Journal of Corpus Linguistics*, vol. 10, no. 1, pp. 85–107, 2005.

[25] A. Cruttenden, *Intonation*. Cambridge: Cambridge Univ Press, 1997.

[26] J. Pierrehumbert and J. Hirschberg, "The meaning of intonational contours in the interpretation of discourse," in *Intentions in Communication*, P. Cohen, J. Morgen, and M. Pollack, Eds. Cambridge: MIT Press, 1990.

[27] C. Gunlogson, "A question of commitment," *Belgian Journal of Linguistics*, vol. 22, no. 1, pp. 101–136, 2008.

[28] M. Steedman, "Information Structure and the Syntax-Phonology Interface," *Linguistic Inquiry*, vol. 31, no. 4, pp. 649–689, September 2000.

[29] A. Anderson, M. Bader, E. Bard, E. Boyle, G. Doherty, S. Garrod, S. Isard, J. Kowtko, J. McAllister, J. Miller *et al.*, "The hcrc map task corpus," *Language and speech*, vol. 34, no. 4, pp. 351–366, 1991.

[30] M. Swerts and R. Geluykens, "Prosody as a marker of information flow in spoken discourse," *Language and speech*, vol. 37, no. 1, pp. 21–43, 1994.

[31] C. Lai, "Prosodic Cues for Backchannels and Short Questions: *Really?*" in *Proceedings of Speech Prosody 2008, Campinas, Brazil, May 2008*, 2008.

[32] K. Evanini and C. Lai, "The importance of optimal parameter setting for pitch extraction," in *Presented at the 2nd PanAmerican/Iberian Meeting on Acoustics, Cancun, Mexico, 15-19 November 2010*, 2010.

[33] G. Kochanski, E. Grabe, J. Coleman, and B. Rosner, "Loudness predicts prominence: Fundamental frequency lends little," *The Journal of the Acoustical Society of America*, vol. 118, p. 1038, 2005.

[34] E. Grabe, G. Kochanski, and J. Coleman, "Connecting intonation labels to mathematical descriptions of fundamental frequency," *Language and Speech*, vol. 50, no. 3, pp. 281–310, 2007.

[35] C. Bartels, *The intonation of English statements and questions: a compositional interpretation*. Routledge, 1999.