Prosodic correlates of directly reported speech: Evidence from conversational speech

Wouter Jansen[†], Michelle L. Gregory[‡], Jason M. Brenier[§]

[†] Department of Linguistics, University of Groningen `w.jansen@let.rug.nl`
[‡] Cognitive and Linguistic Sciences, Brown University `Michelle_L_Gregory@brown.edu`
[§] Department of Linguistics, University of Colorado, Boulder `jbrenier@colorado.edu`

Abstract

This paper investigates the prosodic characteristics of reported speech in the Switchboard corpus. We find that directly reported speech is signalled by a greater overall pitch range than the surrounding narrative material and is typically preceded by intonational phrase boundaries. By contrast, prosody does not seem to distinguish indirectly reported speech from ordinary narrative speech. The implications of these findings for ASR are discussed.

1. Introduction

In this study, we investigate whether prosody distinguishes between directly reported speech and indirectly reported speech, and the implications a prosodic distinction would have for Automatic Speech Recognition (ASR), and speech technology in general.[1] It has been shown that prosodic cues are used to signal discourse functions, particularly in the absence of syntactic or lexical cues ([1], [2]). Especially since the distinction between directly and indirectly reported speech is not always marked by lexical or syntactic means this leads us to expect that the distinction between directly and indirectly reported speech is similarly mirrored by speech prosody. Direct quotes are introduced with verbs of saying, without further lexical marking, as shown in (1). Indirect quotes, however, are often lexically indistinguishable from direct quotes (2), but are optionally marked with a complementizer preceding the reported clause (3). The example in (4), however, demonstrates that even with the presence of a complementizer, the utterance may be ambiguous. All examples are taken from the Switchboard corpus of telephone conversations ([3]).

(1) [T]he wife said "these are not his slacks".

(2) He said he enjoyed it.

(3) Siskel and Ebert said that one of the two liked the whole movie.

(4) [H]e said, "That is why I always give custody to the mama."

Despite their lexical ambiguity, directly reported speech, speech that is generally denoted by quotations in text, and indirectly reported speech have distinct discourse functions. [4]

claim that direct quotes are speech demonstrations, speech that depicts actions much like non-verbal demonstrations. Indirect quotes, on the other hand, function only as descriptions (see also [5]).

Recent research by [6] has suggested that prosody does not distinguish the discourse functions denoted by direct quotes from those denoted by indirect quotes. However, they do not provide any quantitative data and base their conclusions on a relatively small number of cases. The findings by [6] are also contrary to a growing body of research that links the prosodic characteristics of speech to discourse function. For example, aspects of prosody such as pitch range and final lowering have been shown to contribute to discourse interpretation ([1]). Others ([7]) have found that lengthening at prosodic phrase boundaries can predict discourse boundaries. Pitch accent and intonational phrasing have been demonstrated to disambiguate cue phrases ([8]). Additionally, [9] found that prosody aids in automatic dialogue act classification. In the light of the evidence that prosodic structure correlates with discourse function, we test the hypothesis that prosody can distinguish between the discourse functions denoted by direct and indirect quotes in speech.

2. Method

To test whether the distinction between direct and indirect quotes is cued by prosodic features we collected 34 instances of direct quotes (27 produced by female speakers, 7 by male speakers) and 27 instances of indirect quotes (19 female, 8 male) from the Switchboard corpus. All of the speech segments were annotated and analysed using the signal analysis program PRAAT (version 3.9.33). Labelling of accent and boundary features was carried out in accordance with the *POSH* conventions, a transcription system intended for use with natural speech (POSH labeling guide, Ostendorf & Shattuck-Hufnagel, p.c.). We established that all quotes were contained within a single intonational phrase (i.e. a prosodic constituent bounded by breaks with POSH index 3 or 4)

In addition to the intonational phrase containing the quote we also transcribed the intonational phrases flanking it if they are contained in the same utterance. Labelling the preceding and following intonational phrases enabled us to assess the prosodic quotes on a principled basis: the behaviour of many prosodic cues including pitch range has been claimed to be constrained by intonational phrasing (cf. [10]). For each intonational phrase that was transcribed we hand-coded the highest and lowest pitch target in order to obtain pitch measures that are not contaminated by the effects of segmental perturbation or pitch tracker errors caused by creaky voice.

To examine the distinction between direct or indirect quotes

and ordinary narrative speech we extracted 34 random segments of narrative speech (27 female, 7 male) from the files containing the quotes (the Ns were chosen to ensure we had as many narrative segments as quoted segments). The duration of these segments was 2,500 ms, which roughly corresponds to the mean duration of the quotes (2,398 ms). In order to obtain comparable pitch measures for the intonational phrases containing the quotes and the random segments we also hand-coded the overall pitch maxima and minima in the latter.

We measured pitch values in Hertz for the hand labelled pitch targets in the reported speech segments and random narrative segments and the average pitch in Hertz for each intonational phrase in the reported speech segments. In order to limit the influence of segmental pitch perturbation and errors of the pitch tracker caused by excessive creak the average pitch was calculated over all frames with a pitch value between the minima and maxima determined by hand labelling. All measurements were provided by the autocorrelation method pitch tracker included in the PRAAT program (cf. [11]) Pitch maxima and minima were used to calculate overall pitch ranges for (the domains in) all segments. The average pitch within a domain or segment was used as a rough estimator of overall *pitch level* which can vary independently from the pitch range.

3. Results

In order to determine if there is a prosodic distinction between direct and indirect quotes, we focus on three major prosodic characteristics: pitch range, overall pitch level as reflected by the average pitch, and prosodic breaks. We compared the pitch range and mean pitch of the directly reported segments to those of the indirectly reported segments. Secondly, we tested to what extent the pitch range and mean pitch of the quotes were reset relative to the pitch mean and range of the preceding and following intonational phrases. Pitch range and level and the relative size of resets have been observed to reflect syntactic and/or discourse structure in a variety of ways (cf. e.g. [12], [1], [10] [9]).

We also compared whether the two discourse functions could be distinguished by intonational phrase breaks, both preceding and following. Since it has been claimed in the literature (see [10]) that pitch range and level resets often coincide with prosodic breaks established on independent grounds (presence of boundary tones, phrase final lengthening, pausing), it was theoretically possible that significant results for both the pitch variables and prosodic breaks would be circular. We therefore used a logistic regression model to establish the independence of these variables.

3.1. Pitch measurements

Table 1 shows the mean values and standard deviations (in brackets) of pitch range as determined on the basis of the hand coded pitch targets, for the segments produced by female speakers.[2] The means for pitch range indicate that overall range is far greater for direct quotes than indirect quotes and that pitch range does not distinguish between the latter and narrative speech. A one-way ANOVA shows that that there are differences in pitch range among the different types of speech segments, $F(2, 70) =$

---

[2]Because pitch perception and (hence) production is logarithmic rather than linear in the frequency domain, pitch data of the male speakers had to be analysed separately. They seem to show the same patterns as the female data, but the small number of cases does not warrant statistical analysis.

11.64, $p < 0.001$. A Scheffe post-hoc test shows that there are highly significant differences between the intonational phrases containing direct and indirect quotes, $p = 0.002$, and between the direct quotes and the random narrative samples, $p < 0.001$, but not between the random samples an the indirect quotes, ns.

The mean pitch ranges of the intonational phrases preceding those containing the direct and indirect quotes suggest that there also is a considerable difference between direct and indirect quotes in the amount of pitch reset relative to a preceding phrase. On average the overall pitch range on a direct quote is expanded by 47 Hz (sd.=73, N=22) whereas the transition from a preceding phrase to an indirect quote leads to a decrease of 19 Hz (sd.=46, N=17). Paired-samples t-tests indicate that range reset produces a significantly different pitch at the transition into a direct quote, $t(21) = 2.99$, $p = 0.007$, but not at the boundary preceding an indirect quote, $t(16) = -1.78$, $p = 0.094$. Paired-samples t-tests show a (weak) reverse effect for pitch range reset between a quote and the following intonational phrase, $t(11) = -0.17$, $p = 0.87$ for the direct quotes vs. $t(6) = -2.79$, $p= 0.032$ for the indirect quotes, but the number of cases is too small here for the results to be reliable.

Type of segment	prec. IP	quote IP	fol. IP
Direct quote	73 (45, N=22)	119 (66, N=27)	89 (39, N=12)
Indirect quote	85 (41, N=17)	62 (30, N=19)	89 (51, N=7)
Random narrative	n.a.	61 (44, N=27)	n.a.

Table 1: *Mean pitch (standard, deviations) ranges (Hz) for different segment types (female speakers only).*

The pitch range and pitch range reset results indicate that, contrary to the claim by [6], speech demonstrations are in fact prosodically distinct from speech descriptions. Overall pitch range distinguishes direct quotes from indirect quotes and narrative speech while our results show that pitch range is used by speakers to demarcate the onset of direct quotes but not indirect quotes. The data in table 2 indicates that this prosodic distinction between speech demonstrations and speech descriptions does not extend to differences in overall pitch level. Although the numbers suggest that the pitch level of direct quotes is slightly higher than the level of indirect quotes or narrative segments, this is not confirmed by statistical tests. Similarly, there is no statistically significant reset of overall pitch level between direct quotes and the flanking intonational phrases.

Type of segment	prec. IP	quote IP	fol. IP
Direct quote	199 (34, N=22)	205 (41, N=27)	195 (38, N=12)
Indirect quote	191 (34, N=17)	187 (37 N=19)	194 (17, N=7)
Random narrative	n.a.	181 (37, N = 27)	n.a.

Table 2: *Means (standard deviations, number of cases) of overall average pitch (Hz) for different speech segment types (female speakers only).*

3.2. Intonational breaks

The effects we found in the pitch range data are repeated in the occurrence of major prosodic phrase breaks immediately preceding the different types of quotes. The direct quotes in our corpus are 2 times as likely to be preceded by an intonational phrase break (break index 3 or 4) as an indirect quote (see table 3). A chi-square test shows that this difference is highly significant, $\chi^2(1) = 6.77$, $P < 0.001$, and indicates again that prosody directly signals the (onset) of a direct quote. No similar difference was found for the distribution of prosodic breaks at the right edge of the quotes

Type of segment	IP bound.	Weaker bound.	Total
Direct quote	25	9	34
Indirect quote	10	17	27

Table 3: *Prosodic boundaries immediately preceding direct and indirect quotes (Male and female speakers). Break indices 3 and 4 collapsed into single category.*

3.3. Independence of the variables

As noted above, there are claims in the literature that pitch range effects are not independent from the occurrence of prosodic breaks. We used a logistic regression model to test whether pitch range reset and prosodic breaking contribute independently to the signalling of direct vs. indirect quotes. A logistic regression is a statistical model that predicts a dependent variable based on contributions from a number of independent factors ([14]). Using quote type, direct versus indirect, as our independent variable, we tested if pitch reset and breaks independently predict quote type by adding them to the model after pitch range. When pitch reset is added to a model that includes pitch range, which is a significant predictor of quote type, p < 0.0001, is is not significant, ns. However, preceding phrase break is still a good predictor of quote type, even after controlling for pitch range and reset, p<0.01. Thus, we conclude that pitch range and preceding phrase break independently aid in the prosodic distinction between the two quote types.

4. Implications for ASR

The task of converting spoken language into a text stream (ASR) is difficult. While the use of word information alone yields acceptable results in some domain-specific ASR applications, tasks such as automatic transcription and dialogue understanding (ASU) require knowledge of the discourse structure. Knowledge of discourse structure can aid in speech act identification, referent tracking and appropriate punctuation. Prosodic features of speech including duration, F_0, pausing, and energy have been shown to provide important cues to automatic identification of seven speech act types ([13]). Prediction of discourse context and dialog acts was found to substantially improve word recognition accuracy by restricting the number of word possibilities in a given dialog act type.

Our investigation of reported speech in the Switchboard corpus demonstrates that direct quotes are reflected in speech prosody by a greater pitch range, a greater amount of pitch range reset between the intonational phrase containing the quote and the preceding domain, and the tendency for a major break to occur immediately before the quote. By contrast, indirect quotes seem to share the prosody of narrative speech. This knowledge of the prosodic correlates of directly reported speech may further improve automatic discourse context and speech act identification systems. Consequently ASR would also improve if integrated into the same system. The implementation of prosodic information about direct and indirect quotes may also contribute to transcription accuracy by enabling speech recognition systems to insert quotation marks where needed. In a TTS system, the ability to interpret quoted material by assigning appropriate prosodic targets will enable computers to both produce more natural sounding speech and convey important discourse information to users.

The number of direct quotes is relatively small in the Switchboard corpus, 0.5% of the total number of utterances . This low number suggests that the addition of prosodic characteristics for these utterances types would not yield much greater results than word recognition alone. However, quotes are much more frequent in other corpora. Table 4 gives the ratio of the number of quotes over the total number of utterances for the Switchboard, Brown, and Wall Street Journal corpora.

Corpus	Quotes
Switchboard	0.5%
Brown	21%
Wall Street Journal	10%

Table 4: Counts of quotes in 3 different corpora

A Maximum Entropy Parser can parse the Wall Street Journal with 90% accuracy ([15]). However, when punctuation is removed from the text during pre-processing, the performance of the parser drops to 86% (E. Charniak, p.c.). Comma usage and quotation marks have been identified as sources of error. The use of prosodic information may in fact aid parsing in the absence of punctuation, which would eliminate some preprocessing necessary in parsing tasks. The ability to identify the prosodic correlates of punctuation will aid in speech-to-text applications, as well as improve speech parsing.

5. Conclusions

It has been claimed in the literature that reported speech and indirectly reported speech have different discourse functions. The first are can be plausibly construed as speech demonstrations whereas the latter are better treated as descriptions. In this study we have shown that these distinct discourse functions are reflected in the prosodic characteristics of direct and indirect quotes. The former have a greater pitch range and show a greater degree of pitch range reset with respect to the preceding context than the latter. In addition, direct quotes are more likely to be immediately preceded by an intonational phrase break.

While the usefulness of distinguishing the prosodic correlates of the discourse functions of reported speech and indirectly reported speech have not been directly tested in this study, we do provide evidence that such an endeavour is worth pursuing. In addition, these results are the first step in identifying the prosodic correlates of punctuation for applications in ASR and ASU.

6. References

[1] Pierrehumbert, J., and Hirschberg, J., "The meaning of intonational contours in the interpretation of discourse", in Cohen, P., Morgan, J., and Pollack, M. (eds), Intentions in Communication, MIT Press, Cambridge, MA, 1990.

[2] Cutler, A., Dahan, D., and Donselaar, W. van, "Prosody in the Comprehension of Spoken Language: A Literature Review", Language and Speech, 40(2):141-202, 1997.

[3] Godfrey, J., Holliman, E., and McDaniel, J., "SWITCHBOARD: telephone speech corpus for research and development", in Proceedings of ICASSP-92, 517-520, 1992.

[4] Clark, H., and Gerrig, R., "Quotations as demonstrations", Language, 66(4):764-805 ,1990.

[5] Waugh, L., "Reported speech in journalistic discourse: the relation of function and text", Text, 15(1):129-173, 1995.

[6] Klewitz, G., and Couper-Kuhlen, E., "Quote-unquote? the role of prosody in the contextualization of reported speech sequences", Pragmatics, 9(4): 459-485, 1999.

[7] Swerts, M., Collier, R., and Terken, J. "Prosodic predictors of discourse finality in spontaneous monologues", Speech Communication, 15(1-2): 79-90, 1994.

[8] Hirschberg, J., Litman, D., "Empirical studies on the disambiguation of cue phrases", Computational-Linguistics, 19(3):501-530, 1993.

[9] Shriberg, E., Bates, R., Taylor, P., Stolcke, A., Jurafsky, D., Ries, K., Coccaro, N., Martin, R., Meteer, M., and Van Ess-Dykema, C., "Can prosody aid the automatic classification of dialog acts in conversational speech?", Language and Speech, 41(3-4):439-487, 1998.

[10] Ladd, D., Intonational Phonology, Cambridge University Press, Cambridge, 1996.

[11] Boersma, P., "Accurate short-term analysis of the fundamental frequency and harmonics-to-noise ratio of a sampled sound", Proceedings of the Institute of Phonetic Sciences of The University of Amsterdam, 17:97-110, 1993.

[12] Cooper, W., and Sorenson, J., Fundamental Frequency in Sentence Production, Springer, New York, 1981.

[13] Stolcke, A., Coccaro, N., Bates, R., Taylor, P., Van Ess-Dykema, C., Ries, K., Shriberg, E., Jurafsky, D., Martin, R., Meteer, M., "Dialogue act modeling for automatic tagging and recognition of conversational speech", Computational Linguistics, 26(3):339-371, 2000.

[14] Agresti, A, An introduction to categorical data analysis, Wiley. New York, 1996

[15] Charniak, E., "A maximum-entropy-inspired parser", Proceedings of NAACL-2000, 2000.