The vocal intensity of turn-initial cue phrases in dialogue

Anna Hjalmarsson
Department of Speech Music and Hearing, KTH
Stockholm, Sweden
annah@speech.kth.se

Abstract

The present study explores the vocal intensity of turn-initial cue phrases in a corpus of dialogues in Swedish. Cue phrases convey relatively little propositional content, but have several important pragmatic functions. The majority of these entities are frequently occurring monosyllabic words such as “eh”, “mm”, “ja”. Prosodic analysis shows that these words are produced with higher intensity than other turn-initial words are. In light of these results, it is suggested that speakers produce these expressions with high intensity in order to claim the floor. It is further shown that the difference in intensity can be measured as a dynamic inter-speaker relation over the course of a dialogue using the end of the interlocutor’s previous turn as a reference point.

1 Introduction

In dialogue, interlocutors produce speech incrementally and on-line as the dialogue progresses. Articulation can be initiated before the speaker has a complete plan of what to say (Pechmann, 1989). When speaking, processes at all levels (e.g. semantic, syntactic, phonologic and articulatory) work in parallel to render the utterance. This processing strategy is efficient, since the speaker may employ the time devoted to articulating an early part of an utterance to plan the rest.

Speakers often initiate new turns with *cue phrases* – standardized lexical or non-lexical expressions such as “ehm” “okay”, “yeah”, and “but” (c.f. Gravano, 2009). Cue phrases (or *discourse markers*) are linguistic devices used to signal relations between different segments of speech (for an overview see Fraser, 1996). These devices convey relatively little propositional content, but have several important pragmatic functions. For example, these words provide feedback and signal how the upcoming utterance relates to previous context. Another important function is to claim the conversational floor (c.f. Levinson, 1983).

With these fundamental properties of language production in mind, it is proposed that turn-initial cue phrases can be used in spoken dialogue systems to initiate new turns, allowing the system additional time to generate a complete response. This approach was recently explored in a user study with a dialogue system that generates turn-initial cue phrases incrementally (Skantze & Hjalmarsson, in press). Results from this experiment show that an incremental version that used turn-initial cue phrases had shorter response times and was rated as more efficient, more polite and better at indicating when to speak than a non-incremental implementation of the same system. The present study carries on this research, investigating acoustic parameters of turn-initial cue phrases in order to build a dialogue system that sounds convincing intonation wise.

Another aim of this study was to explore if the vocal intensity of the other speaker’s immediately preceding speech can be used as a reference point in order to measure intensity as an inter-speaker relation over the course of a dialogue. Thus, in addition to measuring overall differences in intensity, the relative difference between the first token of a new turn and the last token of the immediately preceding turn was measured. This dynamic approach, if proven feasible, allows spoken dialogue system designers to adjust the system’s vocal intensity on-line in order to accommodate variations in the surrounding acoustic environment.

2 Related work

There are a few examples of research that have manipulated intensity to signal pragmatic functions. For example, Ström & Seneff (2000) increases intensity in order to signal that user

Proceedings of SIGDIAL 2010: the 11th Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 225–228,
The University of Tokyo, September 24-25, 2010. ©2010 Association for Computational Linguistics

barge-ins are disallowed in particular dialogue states. Theoretical support for such manipulations is provided by an early line of research on interruptions in dialogue (Meltzer et al., 1971). Meltzer et al. (1971) propose that the outcome of speech overlaps is affected by prosodic characteristics and show that the greater the increase in amplitude, the greater the likelihood of "interruption success". Moreover, it is show that the success of interruptions, that is who retains the floor, is based on how much higher the intensity of the interruption is compared to the previous speaker's intensity or compared to the speaker's own intensity at the end of that speaker's previous speaker turn.

Measuring inter-speaker relative intensity is further motivated by research that suggests that speakers adjust their vocal intensity online over the course of a dialogue in order to accommodate the surrounding acoustic context. For example, speakers tend to raise their voice unintentionally when background noise increases to enhance their audibility; this is the so-called Lombard effect (Pick et al., 1989). Moreover, speakers adjust intensity based on their conversational partners (Natale, 1975) and the distance to their listeners (Healey et al., 1997).

3 Method

3.1 Data: The DEAL corpus

DEAL is a dialogue system that is currently being developed at the department of Speech, Music and Hearing, KTH (Wik & Hjalmarsson, 2009). The aim of the DEAL dialogue system is to provide conversation training for second language learners of Swedish. The scene of DEAL is set at a flea market where a talking animated persona is the owner of a shop selling used goods.

The dialogue data used as a basis for the data analyzes presented in this paper were human-human dialogues, collected in a recording environment set up to mimic the interaction in the DEAL domain. The dialogue collected were informal, human-human, face-to-face conversation in Swedish. The recordings were made with close talk microphones with six subjects (four male and two female). In total, eight dialogues were collected. Each dialogue was about 15 minutes, making for about two hours of speech in total in the corpus. The dialogues were transcribed orthographically and annotated for entities such as laughter, lip-smacks, breathing and hemming. The transcripts from the dialogues were time-aligned with the speech signal. This was done using forced alignment with subsequent manual verification of the timings. The dialogues were also segmented into *speaker turns*. A speaker turn here is a segment of speech of arbitrary length surrounded by another speaker's vocalization. All together, the dialogues contained 2036 speaker turns.

The corpus was also annotated for cue phrases using 11 functional categories. The definition of cue phrases used for annotation of the DEAL corpus was broad and all types of vocalizations that the speakers use to hold the dialogue together at different communicative levels were included. Cue phrase annotation was designed as a two-fold task: (i) to decide if a word was a cue phrase or not – a binary task, and (ii) to select its functional class according to the annotation scheme. The annotators could see the transcriptions and listen to the recordings while labelling. The kappa coefficient for task (i) was 0.87 ($p<.05$). The kappa coefficient for (ii) was 0.82 ($p<.05$). For a detailed description of the cue phrase categories and their annotation, see (Hjalmarsson, 2008).

3.2 Data analysis

The first word in each turn was extracted and analyzed. Here, a word is all annotated tokens in the corpus except breathing, lip-smacks, and laughter, which are all relevant, but outside the scope of this study. 1137 (57%) words were annotated as some type of cue phrase, and 903 (43 %) were other words. The turn-initial cue phrases were annotated with different cue phrase categories. 587 (28%) turn-initial words were annotated as either RESPONSIVE, RESPONSIVE DISPREFERENCE or RESPONSIVE NEW INFORMATION. The annotation of these was based on the interpretation of the speakers' attitudes, expressing either neutral feedback (RESPONSIVE), non-agreement (RESPONSIVE DISPREFERENCE) or surprise (RESPONSIVE NEW INFORMATION). The RESPONSIVES were most frequently realized as either "ja", "a", and "mm" (Eng: "yeah", "mm").

Furthermore, 189 (9%) of all turn-initial words were annotated as CONNECTIVES. The connective cue phrase categories indicate how the new utterance relates to previous context. For example, these signal whether the upcoming speaker turn is *additive*, *contrastive* or *alternative* to previous context. Examples of these categories are "och" (Eng: "and"), "men" (Eng: "but") and "eller" (Eng: "or"), respectively.

A third category of cue phrases in a turn-initial position was filled pauses (57, 3%). Whereas filled pause may not typically be considered as cue phrases, these elements have similar characteristics. For example, fillers provide important pragmatic information that listeners attend and adjust their behaviour according to. For example, a corpus study of Dutch fillers showed that these tokens highlight discourse structure (Swertz, 1998). Frequently occurring filler words in the corpus were “eh” and “ehm”.

The majority of the turn-initial cue phrases were high frequency monosyllabic words, which are typically not associated with stress, although on listening, they give the impression of being louder than other turn-initial vocalizations. To verify this observation, the intensity in decibel of the first word of each turn was extracted using Snack (www.speech.kth.se/snack). In order to explore the vocal intensity as an inter-speaker relation over the course of the dialogue, the average intensity of the last word of all turns was extracted. The motivation of this approach is to use the previous speaker’s voice intensity as a reference point. Thus, in order to avoid the need for global analysis over speakers and dialogues, only the (un-normalized) difference in intensity between the last word of the immediately preceding turn and the first word of a new turn was calculated.

All turns following a one word only turn from the other speaker were excluded as an approximation to avoid speech following backchannel responses. 300 (33%) of the speaker changes contained overlapping speech. These overlaps were excluded from the data analysis since the recordings were not completely channel-separated and crosstalk could conceivably interfere with the results.

Since the distance between the lips and the microphone was not controlled for during the recordings, the values were first normalized per speaker and dialogue (each value was shifted by the mean value per speaker and dialogue).

4 Results

Figure 1 presents the average normalized intensity for turns initiated with cue phrases and other words.

An independent samples t-test was conducted between the intensity of turns initiated with cue phrases and other turn-initial words. There was a significant difference in intensity between turns initiated with cue phrases (M=3.20 dB, SD=6.99) and turns initiated with other words (M=-4.20 dB, SD=9.98), t(597)=10.55, *p*<.000. This shows that, on average, turns initiated with cue phrases were significantly louder (on average 6 dB) than turns initiated with other words.

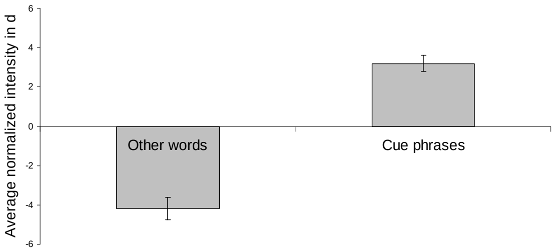

Figure 1 : Average normalized vocal intensity in dB for turn-initial words. Error bars represents the standard error.

In order to explore the vocal intensity as an inter-speaker relation the difference in voice intensity between a new turn and the end of the immediately preceding turn was extracted. The inter-speaker differences in intensity for turn-initial cue phrases and other words are presented in Figure 2.

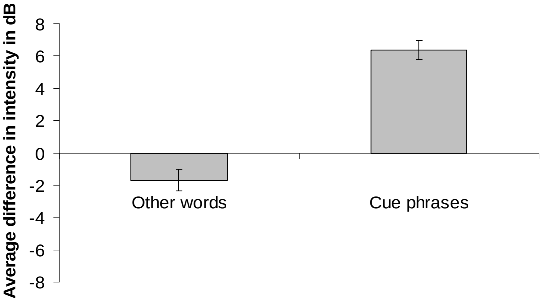

Figure 2 Average difference in intensity (in dB) for turn-initial words. Error bars represents the standard error.

An independent samples t-test was conducted to explore the difference in voice intensity as an inter-speaker relation. There was a significant difference in intensity between turns initiated with cue phrases (M=6.14 dB, SD=11.86) and turns initiated with other words (M=-1.52 dB, SD=13.07); t(595)=7.48, *p*<.000. This shows that the increase in intensity was significantly larger for turns initiated with cue phrases (about 7 dB) than for turns initiated with other words.

5 Discussion

This paper presents analyses of the intensity of turn-initial words. It shown that turns are frequently initiated with cue phrases (about 55% of the turns in the DEAL corpus). The majority of these consist of high frequency monosyllabic words such as “yes”, “mm” and “okay”. The most frequent turn-initial words that were not annotated as cue phrases were “den” (Eng: “it”), “vad” (Eng: “what”), and “jag” (Eng: “I”). Thus, similar to turn-initial cue phrases, this category contains high-frequency monosyllabic words, items that are not typically associated with prosodic stress. Yet, the results show that turn-initial cue phrases are produced with higher intensity than other turn-initial words are. In the light of previous research, which suggests that increased intensity have turn-claiming functions, one can speculate that speakers produce talkspurt-initial cue phrases with increased intensity in order to claim the floor convincingly before having formulated a complete utterance.

It is further argued that turn-initial cue phrases can be used in dialogue systems capable of incremental speech production. Such words can be used to initiate turns once the user has stopped speaking, allowing the system more time to process input without response delays.

Finally, it is suggested that intensity may be better modelled relative to the intensity of the immediately preceding speech rather than in absolute of speaker-normalized terms. Speakers adjust their intensity to the current acoustical environment, and such a dynamic inter-speaker relative model may accommodate the current acoustic context over the course of a dialogue. In support of this approach, the present study shows that the increase in intensity can be calculated dynamically over the dialogue using the end of the previous speaker’s turn as a reference point. Inter-speaker relative measures are also motivated practically. Extracting objective measures of intensity is problematic since contextual factors such as the distance between the microphone and the lips are difficult to control between dialogues and speakers, but the effects are mitigated by dynamic and relative measures. This is not to say that measuring intensity over the course of a single dialogue is trivial. Variation due to for example unforeseen alterations of the distance between the lips and the microphone during the dialogue are still problematic, but it is less of a problem within a session than between different sessions.

Acknowledgments

This research was carried out at Centre for Speech Technology, KTH. Funding was provided by Riksbankens Jubileumsfond (RJ) project P09-0064:1-E Prosody in conversation and the Graduate School for Language Technology (GSLT). Many thanks to Rolf Carlson, Jens Edlund and Joakim Gustafson for valuable comments.

References

Fraser, B. (1996). Pragmatic markers. *Pragmatics, 6*(2), 167-190.

Gravano, A. (2009). *Turn-Taking and Affirmative Cue Words in Task-Oriented Dialogue*. Doctoral dissertation, Columbia University.

Healey, C., Jones, R., & Berky, R. (1997). Effects of perceived listeners on speakers'vocal intensity. *Journal of Voice, 11*(1), 67-73.

Hjalmarsson, A. (2008). Speaking without knowing what to say... or when to end. In *Proceedings of SIGDial 2008*. Columbus, Ohio, USA.

Levinson, S. C. (1983). *Pragmatics.* Cambridge: Cambridge University press.

Meltzer, L., Hayes, D., & Morris, M. (1971). Interruption Outcomes and Vocal Amplitude: Explorations in Social Psychophysics. *Journal of Personality and Social Psychology, 18*(3), 392-402.

Natale, M. (1975). Convergence of mean vocal intensity in dyadic communication as a function of social desirability. *Personality and Social Psychology, 32*(5), 790-804.

Pechmann, T. (1989). Incremental speech production and referential overspecification. *Linguistics, 27,* 89-110.

Pick, H. L. J., Siegel, G. M., Fox, P. W., Garber, S. R., & Kearney, J. K. (1989). Inhibiting the Lombard effect. *JASA, 85*(2), 894-900.

Skantze, G., & Hjalmarsson, A. (in press). Towards Incremental Speech Generation in Dialogue Systems. To be published in *Proceedings of SigDial*. Tokyo, Japan.

Ström, N., & Seneff, S. (2000). Intelligent barge-in in conversational systems. In *Procedings of ICSLP-00*.

Swertz, M. (1998). Filled pauses as markers of discourse structure. *Journal of Pragmatics, 30*(4), 485-496.

Wik, P., & Hjalmarsson, A. (2009). Embodied conversational agents in computer assisted language learning. *Speech communication, 51*(10), 1024-1037.