Pitch similarity in the vicinity of backchannels

Mattias Heldner[1]*, Jens Edlund*[1]*, Julia Hirschberg*[2]

[1] KTH Speech, Music and Hearing, Stockholm, Sweden
[2] Department of Computer Science, Columbia University, New York, NY, USA
mattias@speech.kth.se, edlund@speech.kth.se, julia@cs.columbia.edu

Abstract

Dynamic modeling of spoken dialogue seeks to capture how interlocutors change their speech over the course of a conversation. Much work has focused on how speakers adapt or entrain to different aspects of one another's speaking style. In this paper we focus on local aspects of this adaptation. We investigate the relationship between backchannels and the interlocutor utterances that precede them with respect to pitch. We demonstrate that the pitch of backchannels is more similar to the immediately preceding utterance than non-backchannels. This inter-speaker pitch relationship captures the same distinctions as more cumbersome intra-speaker relations, and supports the intuition that, in terms of pitch, such similarity may be one of the mechanisms by which backchannels are rendered 'unobtrusive'.

Index Terms: backchannels, pitch, interlocutor similarity, inter-speaker features

1. Introduction

In [1], we showed that the tendency for interlocutors to mimic each other's behavior (i.e. *entrainment, priming, accommodation, inter alia* in the literature) can be modeled dynamically over the course of a dialogue. Such dynamically modeled inter-speaker similarity captures the continuous and on-going nature of spoken dialogue and highlights its interactional aspects, whereas much other modeling is more focused on the individual behaviors of the speakers. In this paper we continue to investigate dialogue in terms of its inter-speaker relations. We examine the relationship of a speaker's pitch to that of her interlocutor in short feedback responses compared to other vocalizations.

Short vocalizations such as *mm-hm, okay* and *yeah* can be used to indicate that the speaker producing them is following and understanding, and they encourage the other speaker to proceed [e.g. 2, 3, 4]. These brief utterances are known in the literature as backchannels, continuers, or feedback, and have important communicative and interactive functions. Backchannels are generally described as being somehow produced *in the background*. They are often not taken to constitute a speaking turn or to claim the floor in studies of turn-taking behavior. They may occur in the midst of another speaker's speech without disrupting that speaker [e.g. 5, 6], and they are quieter and shorter than other instances of the same lexical items [e.g. 7, 8-11]. They have also been found, in the corpus studied here, to have higher pitch and to be more likely to bear a rising pitch accent (L+H*) and a high boundary tone (H-H%) than other categories of short vocalizations [12].

In this paper, we focus on how backchannels are rendered *unobtrusive*, beyond previous observations about their voice quality, duration and loudness levels. We examine a previously unstudied aspect of the 'backgrounding' of backchannels in terms of their pitch. We posit that one way of making an utterance *less conspicuous* is to make it *more similar* to the interlocutor's speech. We look for support for this intuition by investigating whether backchannels are more similar to the immediately preceding utterance than non-backchannels with respect to pitch. This could also be described as investigating whether inter-speaker similarity, as far as pitch is concerned, is more pronounced in the vicinity of backchannels than elsewhere in a conversation.

1.1. Backchannels and spoken dialogue systems

A growing field in spoken dialogue system design aims at designing human-like spoken dialogue systems: systems that speak the way people speak to each other, and that encourage their users – their interlocutors, as it were – to behave as when talking to other people [e.g. 13, 14]. A prominent target in this endeavor is to improve the way spoken dialogue systems decide when to speak and when to remain silent. Closely related to this capability is the ability of dialogue systems to understand, appropriately respond to, and produce backchannels. One of the aims of our investigations of backchannels in spontaneous conversation is to improve the way human-like spoken dialogue systems handle such vocalizations. For example, the behavior of a system that encounters speech while it is itself speaking can be greatly improved if the system knows at an early state whether the encountered speech is a backchannel or not. If the user's input indeed represents a backchannel, the system may safely finish what it is saying; if instead the user attempts to claim the turn, the system should cease speaking at its earliest convenience or – if it has something urgent to say – raise its voice and continue speaking [15]. Similarly, a system aiming to behave as humans do should produce backchannels at appropriate places in the dialogue. It thus needs to know not only *when* to produce backchannels, but also *how* they should be produced and what responses they are likely to elicit. (See [16] for a description of how a system's use of backchannels affects user behavior.)

1.2. Inter-speaker relative descriptions

As indicated above, backchannels have been described as having prosodic characteristics which differentiate them from other vocalizations. Some of these seem to be intrinsically relative. *Quiet*, for example, like most prosodic characteristics, makes sense only in relation to something else – to some model of loudness. A general model would capture how loudly a speaker speaks on average. This model can be acquired once, and any speech can be compared with it to give a general idea of whether it is loud or not. This generality comes at a price, however, and a static model would fail to filter out variation in loudness for other reasons. For example, we have Lombard effects caused by variable background noise and variations in the theme and intensity of the dialogue. A more specific model might be acquired by tracking a speaker's loudness over a conversation, permitting one to measure the speaker's relative distance from her own recent production, which normalizes out some of the variation. As a final example, if we measure

the distance between the current speaker and her interlocutor, we obtain a dynamic and current measure that is less sensitive to influences affecting both speakers simultaneously. Examining other features, such as pitch, in this more dynamic way may lead to similar robustness gains.

2. Method

2.1. Columbia Games Corpus

The data used in this work is drawn from the Columbia Games Corpus, a collection of spontaneous task-oriented dialogues by native speakers of Standard American English, and its associated annotations. This corpus contains recordings made using close-talking microphones, with speakers recorded on separate channels, 16 bit/48 kHz, in a sound-proof booth. Speakers were asked to play two types of collaborative computer games that required verbal communication. The speakers did not have eye contact. There were 13 subjects (7 males and 6 females) and they formed 12 different speaker pairs. Eleven of the subjects spoke with two different partners in two separate sessions. The recording sessions lasted on average 45 minutes, and the total duration of the corpus is 9 hours 8 minutes.

The corpus has been orthographically transcribed and manually annotated for a number of phenomena. For the present study, we have primarily used the labeling of single *affirmative cue words* (i.e. lexical items potentially indicating agreement such as *alright, gotcha, huh, mm-hm, okay, right, uh-huh, yeah, yep, yes, yup*) with their communicative function, by three trained annotators, and the labeling of turn-exchanges, by two trained annotators. One function labeled for affirmative cue words was 'backchannel'; others were *affirmation/agreement, cue phrase beginning discourse segment, cue phrase ending discourse segment, pivot beginning* and *pivot ending*: variants of the previous categories in which both cue phrase and affirmation/agreement functions were present, *literal modifiers, return from a previous task, checks* and *stalls*. Turn exchanges were labeled by first identifying *Interpausal Units* (IPUs), maximal sequences of words surrounded by silence longer than 50 ms [cf. talkspurts in 17]. A turn was defined as a maximal sequence of IPUs from a single speaker, so that between any two adjacent IPUs there is no speech from the interlocutor [cf. talkspurts in 18].

All turn transitions in the corpus were classified using a labeling scheme adapted from [19] that identifies, inter alia, *smooth switches (*S*)* — transitions from speaker A to speaker B such that (i) A manages to complete her utterance, and (ii) no overlapping speech occurs between the two conversational turns; *pause interruptions (*PI*)*, defined as cases similar to smooth switches except that A does ***not*** complete her utterance; and *backchannels (*BC*)*, defined as an utterance produced a "response to another speaker's utterance that indicates only *I'm still here / I hear you and please continue*", with no attempt to take the turn. Speech from A following backchannels from B was labeled separately as X2. All continuations from one IPU to the next IPU within the same turn were automatically labeled as HOLD transitions. See http://www.cs.columbia.edu/speech/games-corpus/ for further details and annotation manuals.

2.2. Data

For the present study, we examined transitions involving a speaker change in the Columbia Games Corpus. Speaker HOLDs were excluded, as they are not directly relevant for inter-speaker relations. Transitions with overlapping speech were excluded to guarantee that all pitch analyses remained untainted by crosstalk. Our primary interest was the comparison of backchannels (BC), smooth switches (S), and pause interruptions (PI), which were all included. We also included speech following backchannels (X2) for completeness, as this is the only remaining category of speaker change in silence, and as it might provide insights as to what happens "on the other side" of a backchannel.

In subsequent statistical analyses, we collapsed the data for smooth switches and pause interruptions (S+PI), as these categories are similar and as there are relatively few pause interruptions. Thus, we contrasted backchannels with smooth switches plus pause interruptions, and utterances following backchannels. In addition, we contrasted backchannels with a collapsed category including all other single affirmative cue words (AFFCUE). The backchannel category in both comparisons was identical, while the other discourse functions of affirmative cue words comprised a subset of the smooth switches plus pause interruptions category. The latter comparison was motivated primarily by the shared vocabulary of backchannels and other affirmative cue words.

2.3. Prosodic features

We used a Praat script, the Prosogram v2.6 (http://bach.arts.kuleuven.be/pmertens/prosogram/) to extract the fundamental frequency (F0) data upon which all pitch features were based. The Prosogram provides a perceptually motivated stylization of the F0 contours in voiced portions of local intensity maxima. In other words, the Prosogram has the desirable property of providing reduced and stylized descriptions of pitch patterns in intervals approximating syllable nuclei [e.g. 20]. The following parameter settings were used: a frame period of 0.01 s; an automatic segmentation using the intensity of the band-pass filtered signal; a glissando threshold $G=0,16/T^2$ (where T is the duration of the analyzed segment); and a differential glissando threshold DG of 20.

All pitch features were calculated from stylized semitone transformed F0 values from all dialogue segments containing a single speaker. Segments with overlapping speech were excluded throughout to avoid the risk of crosstalk contaminating the pitch values. Speaker means and standard deviations were calculated for all speaker/session pairs. These statistics were used for all speaker based normalizations.

To describe the transitions, we used the mean pitch over the *last* 500 ms preceding (and including) the last voiced frame before the transition, and the mean pitch over the *first* 500 ms following (and including) the first voiced frame after the transition. Instances where less than 50 ms of either 500 ms interval was voiced were excluded from further analyses. For comparison, we extracted traditional, individual pitch means (based on the current speaker only) by (i) calculating the distance from the speaker's mean over the current session (OWN). We also calculated relative pitch distances between speakers across the transitions based on (ii) raw pitch (RAWREL), (iii) mean normalized pitch (MEANREL), and (iv) z-score normalized pitch (ZREL). A hybrid measure which can be seen as part individual and part relative was also included: the distance between the current speaker's raw pitch and the previous speaker's overall pitch mean (OTHER).

3. Results

3.1. Inter-speaker distances

The analysis of pitch distance between consecutive utterances revealed that backchannels are different from other kinds of utterances, and that they are more similar to the preceding

utterance with respect to pitch than to e.g. the prior speaker's continuation after the backchannel. An ANOVA with pitch distance in semitones (RAWREL) as dependent variable and turn exchange type (three levels: BC vs. S+PI vs. X2) as independent variable showed a significant main effect *F(2,4165)=37; p<.01*. A Tukey HSD post hoc test showed that all three levels of the independent variable were significantly different. Thus, backchannels (mean 0.7 ST) were significantly ***closer*** to the preceding utterance; smooth switches plus pause interruptions were significantly ***higher*** than the preceding utterance (mean 1.8 ST); and utterances following a backchannel (i.e. X2) were significantly ***lower*** than the preceding backchannel (mean -1.7 ST), cf. Figure 1.

Similarly, backchannels were on average 1.3 ST ***closer*** to the preceding utterance than the AFFCUE category (other discourse functions of affirmative cue words; mean 2.0 ST). This difference, too, was significant in an ANOVA: *F(1,1822)=10; p<.01*.

The same analyses using MEANREL (i.e. mean normalized pitch) and ZREL (i.e. z-score normalized pitch) produced qualitatively identical results. Therefore, these results have been omitted from the presentation.

3.2. Model based pitch

The analyses of pitch distance between utterance and session-wide models of speaker pitch range generally showed that backchannels were higher in pitch than other types of utterances. The main effect of turn exchange type (BC vs. S+PI vs. X2) was significant both in the ANOVA with distance from the mean of the same speaker's pitch range (i.e. OWN) as dependent variable: *F(2,4165)=27; p<.01;* and in the ANOVA with distance from the mean in the previous speaker's pitch range (i.e. OTHER) as dependent variable: *F(2,4165)=6; p<.01*. Tukey HSD post hoc tests showed that all three levels were significantly different in the OWN analysis, while only backchannels were significantly different from the other levels in the OTHER analysis. Thus, backchannels (mean 1.7 ST) were significantly higher in pitch than both S+PI (mean 0.7 ST) and X2 (mean 0.2 ST) in the OWN analysis, cf. Figure 2. In the OTHER analysis, backchannels were significantly higher (mean 1.6 ST) than both S+PI (mean 0.6 ST) and X2 (mean 0.1 ST), while S+PI and X2 were not significantly different.

Similarly, backchannels were on average 0.8 ST higher than the AFFCUE category in the OWN analysis (mean 0.9 ST), and 0.9 ST higher than the AFFCUE category in the OTHER analysis (mean 0.7 ST). Both these mean differences were significant: *F(1,1822)=18; p<.01* (OWN); and *F(1,1822)=6; p<.05* (OTHER).

4. Discussion

The results show that a dynamic inter-speaker pitch relation – the difference between the pitch of the speech immediately preceding a speaker change and the pitch speech following immediately after the change – captures the same distinctions as the more cumbersome and less direct intra-speaker relation between a speaker's mean pitch and current pitch. Each difference observed in the OWN analysis (significant differences between BC, S+PI, and X2; significant differences between BC and AFFCUE) was also observed and significant in RAWREL (as well as MEANREL and ZREL). We argue that all of the statistical mean differences found are large enough to have perceptual relevance.

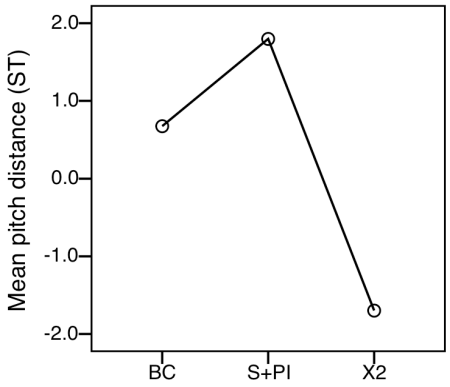

Figure 1: *Pitch distance relative to previous utterance in semitones (RAWREL) for backchannels (BC), smooth switches plus pause interruptions (S+PI), and utterances following a backchannel (X2).*

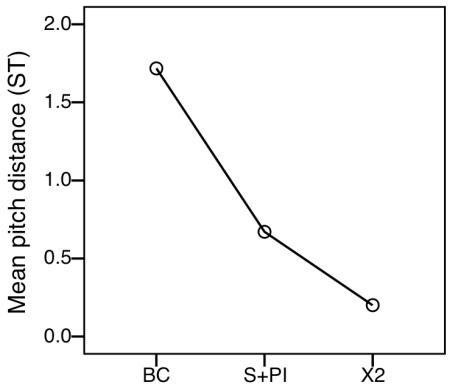

Figure 2: *Pitch distance relative to own pitch mean in semitones (OWN) for backchannels (BC), smooth switches plus pause interruptions (S+PI), and utterances following a backchannel (X2).*

The finding is of interest for spoken dialogue system design, since keeping track of the immediately preceding pitch (from system or from user) is direct and dynamic and eliminates the need to keep track of speaker statistics or to build static speaker models of for example pitch range. Inter-speaker relations are also potentially more robust against variation (e.g. ambient noise, engagement, dialogue type).

The results also show that backchannels are indeed more similar to the preceding utterance in pitch than other utterances, or, put differently, that inter-speaker similarity is more pronounced in backchannels than elsewhere. As far as we know, this is a novel contribution, and it lends support to the intuition that making an utterance more similar to another speaker's speech with respect to pitch is a means of making it less conspicuous.

We also note that backchannels were produced on average 1.7 ST above a speaker's mean, which is consistent with [12], while at the same time being on average only 0.7 ST above the previous speaker (compare Figures 1 and 2). Thus, backchannels become more similar by meeting a raised pitch in the preceding utterance. This suggests that the speech preceding backchannels is also higher than the speaker's mean, and strengthens the idea that the speaker of the backchannel goes out of her way to match the preceding

speakers final pitch in these transitions, rather than just meeting by coincidence.

The pattern remains the same when we compare backchannels to other affirmative cue words with a near-identical vocabulary: backchannels are relatively closer to the preceding utterance, while being higher compared to their speaker's mean. Based on these findings, we speculate that similarity to the preceding utterance is a prosodic characteristic of backchannels.

We also note that X2 – utterances following a backchannel – are produced on average at a pitch which is very close (0.1 ST) to the speaker's mean, while at the same time being on average 1.7 ST lower than the preceding backchannel, suggesting a pattern of a first utterance ending high in a speaker's range, followed by a backchannel equally high, followed by more speech from the first speaker starting at this speaker's mean pitch.

We included for completeness the OTHER analysis, which can be said to combine inter-speaker relations with static modeling. It produces results that are very similar to the OWN analyses, but one significant distinction is lost.

5. Conclusions and future work

We have shown that the pitch at the beginning of a backchannel is similar to the pitch at the end of the utterance that precedes it. This relationship appears particular to backchannels and the utterances preceding them, when we compare the pitch distances across other types of non-overlapping turn exchanges, they are without exception larger than in the backchannels. This is also an indication that the tendency towards interlocutor similarity with respect to pitch is stronger in backchannels than in other types of utterances. We view this study as a starting point for attempts at a more dynamic modeling of dialogue, with a stronger focus on the relations between the speakers.

Next steps include tuning, testing, as well as adding other inter-speaker features (e.g. loudness relations and other inter-speaker pitch features), for example by including them in a dialogue act classification task. Another obvious extension is to test the parameters on materials with greater variability – for example variation in the engagement of the interlocutors or in the ambient noise level – to quantify their robustness.

6. Acknowledgements

This research was carried out at the Department of Computer Science, Columbia University, New York. Funding was provided by the Riksbankens Jubileumsfond (RJ) project P09-0064:1-E *Prosody in conversation* and by NSF IIS-0307905.

7. References

[1] Edlund, J., Heldner, M., and Hirschberg, J., "Pause and gap length in face-to-face interaction", In *Proceedings of Interspeech 2009*, 2779-2782, 2009.

[2] Allwood, J., Nivre, J., and Ahlsén, E., "On the semantics and pragmatics of linguistic feedback", *Journal of Semantics,* 9:1-26, 1992.

[3] Schegloff, E., "Discourse as an interactional achievement: Some uses of 'uh huh' and other things that come between sentences", In D. Tannen [Ed.], *Analyzing Discourse: Text and Talk*, 71-93, Georgetown University Press, 1982.

[4] Clark, H. H., *Using language*, Cambridge University Press, 1996.

[5] Yngve, V. H., "On getting a word in edgewise", In *Papers from the Sixth Regional Meeting Chicago Linguistic Society*, 567-578, Chicago Linguistic Society, 1970.

[6] Gravano, A., "Backchannel-inviting cues in task-oriented dialogue", In *Proceedings Interspeech 2009*, 1019-1022, 2009.

[7] Caspers, J., "Local speech melody as a limiting factor in the turn-taking system in Dutch", *Journal of Phonetics,* 31:251-276, 2003.

[8] Dhillon, R., Bhagat, S., Carvey, H., and Shriberg, E., "Meeting recorder project: Dialog act labeling guide," ICSI Technical Report TR-04-002, 2004.

[9] Ward, N., "Pragmatic functions of prosodic features in non-lexical utterances", In *Proceedings of Speech Prosody 2004*, 325-328, 2004.

[10] Shriberg, E.*, et al.*, "Can prosody aid in the automatic classification of dialog acts in conversational speech", *Language and Speech,* 41:439-487, 1998.

[11] Koiso, H., Horiuchi, Y., Tutiya, S., Ichikawa, A., and Den, Y., "An analysis of turn-taking and backchannels based on prosodic and syntactic features in Japanese map task dialogs", *Language and Speech,* 41:295-321, 1998.

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

[13] Cassell, J., "Body language: Lessons from the near-human", In J. Riskin [Ed.], *Genesis Redux: Essays in the History and Philosophy of Artificial Life*, 346-374, The University of Chicago Press, 2007.

[14] Edlund, J., Gustafson, J., Heldner, M., and Hjalmarsson, A., "Towards human-like spoken dialogue systems", *Speech Communication,* 50:630-645, 2008.

[15] Ström, N. and Seneff, S., "Intelligent barge-in in Conversational Systems", In *Proceedings ICSLP 2000*, 2000.

[16] Gustafson, J., Heldner, M., and Edlund, J., "Potential benefits of human-like dialogue behaviour in the call routing domain", In *Perception in Multimodal Dialogue Systems*, 240-251, Springer, 2008.

[17] Brady, P. T., "A statistical analysis of on-off patterns in 16 conversations", *The Bell System Technical Journal,* 47:73-91, 1968.

[18] Norwine, A. C. and Murphy, O. J., "Characteristic time intervals in telephonic conversation", *The Bell System Technical Journal,* 17:281-291, 1938.

[19] Beattie, G. W., "Turn-taking and interruption in political interviews: Margaret Thatcher and Jim Callaghan compared and contrasted", *Semiotica,* 39:93-114, 1982.

[20] Mertens, P., "The Prosogram : Semi-automatic transcription of prosody based on a tonal perception model", In B. Bel and I. Marlien [Eds.], *Speech Prosody 2004, International Conference (SP-2004)*, 549-552, 2004.