Duration and pitch in perception of turn transition by Swedish and English listeners

Margaret Zellers
Department of Speech, Music & Hearing, KTH, Stockholm, Sweden
zellers@kth.se

Abstract

Turn transition is often predictable, as evidenced by the relative ease with which speakers follow one another without large silences between turns. The current study investigates prosodic turn-taking cues in perception in Swedish and English. While Swedish listeners prefer duration as a cue for turn transition, English listeners prefer pitch cues. This difference between languages may be due to the relatively heavier meaning-bearing role for pitch in Swedish.

Introduction

The apparent ease with which transition from one speaker to another occurs in conversation has been widely acknowledged ever since Sacks, Schegloff & Jefferson's (1974) seminal paper, although some more recent work has indicated that these transitions are not as smooth as those authors originally proposed. Heldner & Edlund (2010) found a large degree of variation in timing of the onsets of new speakers' turns, with the greatest number of turn transitions across three corpora having a "just-noticeable" silence with a duration of around 200ms. Heldner (2011) showed further that overlaps and gaps at turn transitions must be at least 120ms long in order to be perceptible to listeners as overlaps or gaps. Heldner & Edlund (2010) indicate that in their data, at least 41% of cases had a long enough gap that the next speaker could potentially use phonetic information from even the very end of the prior turn as a cue to turn transition.

Prosodic turn transition cues

In English, intonational patterns at prosodic boundaries have been associated with cuing turn hold or transition. Local, Kelly & Wells (1986) note that high pitch rises or low falls generally occur at speaker transition in Tyneside English, alongside other phonetic features such as slowing and vowel centralization. This intonational finding has recently been replicated by Gravano & Hirschberg (2009, 2011) in a corpus of American English. Ford, Fox & Thompson (1996) and Schegloff (1998) also point out that intonation can influence whether complete syntactic structures are interpreted as turn ends in English conversation, countering arguments from e.g. de Ruiter et al. (2006) that intonation does not play a role in the predictability of turn ends.

The case of prosodic cues at turn boundaries in Swedish appears to be somewhat more complicated. Hjalmarsson (2011) found that Swedish listeners associated "flat" intonation with turn-holding, while "falling" intonation was associated with turn-yielding. This is consistent with findings by Edlund & Heldner (2005), who also found that rising intonation patterns were not consistently associated with either turn-holding or turn-yielding. Although Hjalmarsson (2011) did not find that listeners made consistent predictions of turn-transition on the basis of final lengthening, Hjalmarsson & Laskowski (2011) showed that including final lengthening in an automatic model improved prediction of speaker change at pauses, with increased final lengthening before the pause being associated with turn hold.

Perception experiments

A set of perception experiments were conducted in which listeners in either Swedish or English made judgments about prosodic features related to turn transition.

Methodology

Spoken sentences were taken from corpora of conversational Swedish (DEAL corpus, Hjalmarsson et al., 2007) and British English (unpublished corpus, Cambridge University). The conversational turns used for the experiment were chosen on the basis of several criteria. First, they were syntactically complete, but with a declarative sentence form (i.e. not interrogative or imperative). This meant that the turns were syntactically/semantically ambiguous as to turn transition. The final word of each turn was a content word, with stress on the penultimate or antepenultimate syllable. In the Swedish turns, the final word was focally accented (i.e. an additional high pitch peak followed the word-accent LH tones; cf. Bruce, 1998; Gårding 1989). In the English turns, the final word was pitch-accented, always with an H* tone. The pitch contours in all turns in both languages ended with a final fall to low (i.e. L%). There were four base turns each for the Swedish and the English experiments.

Resynthesis of turn prosody

Modifications of duration and pitch characteristics of the turns were carried out using PSOLA resynthesis in Praat (Boersma & Weenink, 2013). The final unstressed syllable(s) of each turn had their duration modified, so that segments in the final rhyme had a duration of either 0.1 sec (Short condition) or 0.15 sec (Long condition).

The pitch contours of the turns were then modified in two ways, illustrated in Figure 1. First, the height of the final pitch peak (i.e. the focus tone peak in Swedish and the pitch accent peak in English) was modified to a level of either 3, 5, or (in Long stimuli) 8 semitones (st) above the speaker's baseline pitch. These modifications will be referred to hereafter as the Peak modifications.

For each peak height, a set of modifications of the final pitch fall were made, with the fall ending at the speaker's baseline, or else 2, 4, or 6 st above the speaker's baseline. In all cases the basic shape of a pitch fall was retained, so items with Peak height 3 st only had falls to baseline and 2 st above the baseline, while items with Peak height 8st had falls to 0, 2, 4, and 6 st. The modifications of the ends of the pitch contours will be referred to hereafter as the Truncation modifications. A schematic of the pitch manipulations is given in Figure 1.

After the resynthesis of the turns was completed, five native speakers of Swedish and four native speakers of English gave naturalness ratings for each of the stimuli. Stimuli which did not attain a majority rating as reasonably natural were not included in the final experiment. 50 of the 56 original Swedish stimuli remained, while 55 of the 56 English stimuli were rated as acceptable. This allowed for 101 usable pairs in the Swedish stimuli. The English stimuli used were matched to the Swedish stimuli, with sentences paired on the basis of highest average acceptability.

Figure 1. Schematic of pitch manipulations made in the experiment stimuli. Left: Peak variations. Right: Trunc variations (for a stimulus with Peak height 5st).

Experiment procedure

In each trial, participants heard two versions of one of the base sentences. The two versions differed in one, two, or all three of the prosodic manipula-

tions described earlier. Half of the participants were asked to choose which version of the sentence the speaker would say if s/he had more to say and was going to continue speaking (Hold condition), and the other half to choose which version the speaker would say if s/he was done speaking and ready for someone else to talk (Change condition). Participants had the option to relisten to the pairs, but most did not do so after the first few trials. Participants also reported for each response whether they were relatively sure or relatively unsure about their selection.

Thirty-two native speakers of Swedish, all resident in the Stockholm area, participated in the Swedish version of the experiment. Twenty-four native speakers of British English, all resident in Cambridge, UK, participated in the English version of the experiment.

Results

Swedish

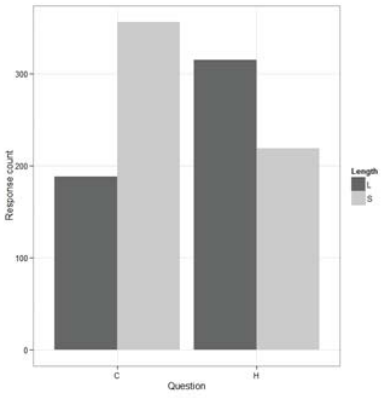

Figure 2. Length of chosen stimuli in Change (C) and Hold (H) conditions for Swedish listeners. ($\chi 2(1, N=1978)= 63.6423, p<0.001$).

The Swedish listeners preferred to use duration cues when they were available. In turns where there was a difference in the Length condition between the two stimuli, listeners in the Hold condition preferred the Long stimuli, while listeners in the Change condition preferred the Short stimuli, as shown in Figure 2.

When there was no duration difference between the two stimuli presented, the Swedish listeners tended to prefer stimuli with high Peak values in the Hold condition and low Peak values in the Change condition. Although the most commonly chosen stimuli in the Hold condition also had high Truncation values, and the most commonly chosen stimuli in the Change condition had low Truncation values, these differences did not attain statistical significance when Peak height was included in the statistical model. It is also important to note that Peak height did not override Length in listeners' judgments about appropriateness for turn transition.

English

The English listeners, in contrast to the Swedish listeners, responded very strongly to both Peak and Truncation variations, with high Peaks and high Truncation levels being associated with Hold, and low Peaks and Truncation with Change. Length variations did not appear to play a role in the English listeners' judgments about which turn versions were more appropriate for Hold or Change. The Peak and Truncation levels preferred by the English listeners are shown in Figure 3 overleaf.

Discussion

The evidence from these two perception experiments demonstrates that listeners can judge whether a turn is ending or not at least in part on the basis of the prosodic form of that turn. Since not all syntactically/semantically complete productions necessarily lead to turn transition, the prosodic form of these turns can be a valuable tool for the conversational participant who is determining whether and when to begin a new turn.

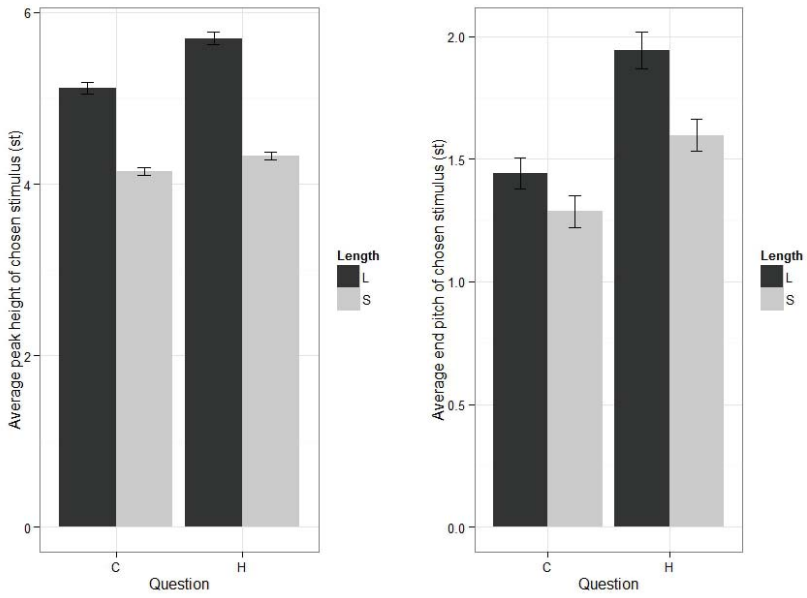

Figure 3. Height of pitch Peaks (left) and Truncation (right) for English listeners in Change (C) and Hold (H) conditions. Length of stimuli did not have a statistically significant effect on listeners' judgments. Peaks: $F(1, 2422)=33.16$, $p<0.001$. Trunc: $F(1,2422)=36$, $p<0.001$.

The prosodic cues reported here should be seen as probabilistic rather than definitive, for several reasons. First, although listeners were sensitive to the cues, they were not forced into particular interpretations; that is, a Short turn (in Swedish) or a turn with a low pitch Peak (in English) did not guarantee that the listener would consider that turn as over; instead, the results represent preferences over many repetitions of similar stimuli. Second, the cues occurred only at the very ends of turns – within the last 500ms or so. In the current experiment, the turns were always followed by silence, so there was no question that the participants had time to hear the prosodic cues and respond to them. However, turns beginning less than about 150ms after the offset of the previous turn are likely not to reflect responses to turn-final prosodic detail. Average reaction time to auditory stimuli in a variety of experiments tends to be around 140-160ms (see review in Kosinski 2013), and turns which begin in overlap do not take into account the previous speaker's intention, whether or not that speaker then chooses to produce a prosodic cue indicating their intention about what may follow their turn (i.e. as a result of earlier planning that is not modified on the basis of the overlapping speech).

Cross-linguistic differences

While Swedish listeners preferred duration variation as a cue to turn transition, English listeners preferred pitch variation. Pitch has regularly been reported as being strongly linked with turn transition in English (Local et al., 1986; Gravano & Hirschberg 2009, 2001 *inter alia*), while its association with turn transition in Swedish has been more ambiguous (Edlund & Heldner, 2005; Hjalmarsson, 2011). Similarly, duration variations with longer duration being associated with turn hold, have been reported for Swedish production data (Hjalmarsson & Laskowski, 2011;

Zellers, submitted) as well as for English (Gravano & Hirschberg, 2009, 2011). If these kinds of variations are available in both languages'' productions, why are listeners with different native languages so different in their preference for turn transition cues?

One possible explanation for this phenomenon depends on the intonational phonology of the two languages in question. Central Swedish has a complex word-accent system (Bruce, 1998; Gårding, 1989) in which most or all content words bear a potentially contrastive word accent. Focus is additionally marked with an additional high (H) tone following the focused content word. English, in comparison, has a relatively sparse tonal specification. Content words may bear a pitch accent but need not, especially if they refer to already-given information, and focus is typically marked by the placement of the nuclear pitch accent, and possibly a following phrase boundary, as well as phonetic modification of that accent to increase its prominence (cf. e.g. Gussenhoven, 2004).

If we assume a cross-linguistics preference for pitch as a signaling tool – perhaps because it may be more salient than other types of prosodic variation – then we can argue that English, with its relatively sparse use of pitch variation in the intonational structure, has "space" left for pitch variation to be used as a turn-transition cue. Swedish, on the other hand, may already be saturated when it comes to pitch as a meaning-bearing tool. Thus, the default turn transition cue must be something else – e.g. duration, as found by the current perception experiment.

A short further investigation of pitch variation in Swedish production data lends further support for this proposal. Turns or turn segments in 10 minutes of spontaneous Swedish conversation from the DEAL corpus were investigated with regards to their prosodic characteristics. In turns with similar form to the Hold and Change turns used in the perception study above (i.e. syntactically complete with indicative structure, and with the final word being a focused content word), duration was found to vary consistently with the listeners' preferences in the perception experiment: that is, potential boundaries followed by a short silence then continuing talk by the same speaker had relatively long final syllables, while potential boundaries followed by a short silence and then talk from another speaker had relatively short final syllables. However, the pitch characteristics of these turns (height of final pitch peak, height of final L%, or height of low pitch preceding final pitch peak) did not vary significantly in relation to the turn transition structure. The only source of statistically significant variation in these pitch data was whether the content word bore Accent 1 (in which case the first L tone in the final LHL sequence is considered to be associated with the stressed syllable) or Accent 2 (in which case the first L tone is considered to be a trailing tone after the starred H tone). In other words, the only statistically significant pitch variation is directly tied to the intonational phonology.

Conclusions

The current perception study has demonstrated that prosodic cues are relevant to listeners as part of their decision about the turn-taking process in conversation. The study gives evidence for language-specific variation, potentially influenced by the phonological system of the language in question. Further research in other, more varied languages, will give further insight into the degree of overlapping influence of the phonological system on the turn-taking system or vice versa.

Acknowledgements

I am very grateful to David House, Anna Hjalmarsson, Jana Götze, Mattias Heldner, Myra Öberg, Niklas Vanhainen, Brechtje Post, Francis Nolan, Calbert Graham, and Elaine Schmidt for assistance with various

aspects, and in particular to Sarah Hawkins and Richard Ogden for use of their unpublished corpus of English conversational data. This research was supported by the postdoctoral grant "Perception of prosody in linguistic contexts" (VR-435-2011-6871) from the Swedish Research Council (Vetenskapsrådet).

References

Boersma, P. & Weenink, D. (2013). Praat: doing phonetics by computer [Computer program]. Available http://www.praat.org/

Bruce, G. (1977) Swedish word accents in sentence perspective. Lund: Gleerup.

De Ruiter, J.P., Mitterer, H. & Enfield, N.J. (2006) Projecting the end of a speaker's turn: a cognitive cornerstone of conversation. Language 82(3): 515-535.

Edlund, J. & Heldner, M. (2005) Exploring prosody in interaction control. Phonetica 62(2-4): 215-226.

Ford, C.E., Fox, B.A., & Thompson, S.A. (1996) Practices in the construction of turns: the "TCU" revisited. Pragmatics 6(3): 427-454.

Gravano, A. & Hirschberg, J. (2009) Turn-yielding cues in task-oriented dialogue. Proceedings of SIGDIAL 2009, Queen Mary University of London, UK, 253-261.

Gravano, A. & Hirschberg, J. (2011) Turn-taking cues in task-oriented dialogue. Computer Speech and Language 25: 601-634.

Gussenhoven, C. (2004) *The phonology of tone and intonation*. Cambridge, UK: Cambridge University Press.

Gårding, E. (1989) Intonation in Swedish. Lund University Department of Linguistics Working Papers 35: 63-88.

Heldner, M. (2011) Detection thresholds for gaps, overlaps, and no-gap-no-overlaps. Journal of the Acoustical Society of America 130(1): 508-513.

Heldner, M. & Edlund, J. (2010) Pauses, gaps and overlaps in conversation. Journal of Phonetics 38: 555-568.

Hjalmarsson, A. (2011) The additive effect of turn-taking cues in human and synthetic voice. Speech Communication 53: 23-35.

Hjalmarsson, A. & Laskowski, K. (2011) Measuring final lengthening for speaker-change prediction. Proceedings of 12th Interspeech, Florence, Italy.

Hjalmarsson, A., Wik, P. & Brusk, J. (2007) Dealing with DEAL: a dialogue system for conversation training. Proceedings of SIGDIAL, Antwerp, Belgium, 132–135.

Kosinski, R.J. (2013) A literature review on reaction time. [online] Available http://biae.clemson.edu/bpc/bp/ Lab/110/reaction.htm.

Local, J.K., Kelly, J. & Wells, W.H.G. (1986) Towards a phonology for conversation: turn-taking in Tyneside English. Journal of Linguistics 22: 411-437.

Sacks, H., Schegloff, E.A. & Jefferson, G. (1974) A simplest systematics for the organisation of turn-taking for conversation. Language 50(4): 696-735.

Zellers, M. (2013) Pitch and lengthening as cues to turn transition in Swedish. Proceedings of 14th Interspeech, Lyon, France.

Zellers, M. (submitted) Prosodic variation for turn transition in Swedish.