Intonation as a cue to turn management in telephone and face-to-face interactions

Miguel Oliveira, Jr.* and Tiago Freitas**

*School of Psychology, St. Andrews University, **ILTEC, Lisbon
miguel.oliveira@st-andrews.ac.uk, taf@iltec.pt

Abstract
Melodic properties of turn boundaries in face-to-face and non-face-to-face interactions are analyzed from a production and a perception viewpoint. Results show that although listeners seem to be able to predict turn management in telephone speech when some conditions are met, melodic information by itself is for the most part insufficient to indicate turn transition to conversationalists.

1. Introduction
The exchange of speaker participation is commonly referred to as “turn-taking” [1] and [2]. Instead of being a haphazard phenomenon, turn-taking is regulated by a series of rules which are often linguistic in nature. Phenomena such as the completion of a grammatical subject-predicate string, and/or the use of a falling pitch at the end of an utterance and open configuration of the vocal tract [3], serve as cues to indicate the end of a turn, or to invite the hearer to take a turn. Conversely, a speaker may maintain his or her turn by increasing volume and speech rate, or by resorting to vocal lengthening. Moreover, it has been observed that in typical face-to-face interactions non-linguistic, physical behaviors are of great importance in regulating turn-taking. As discussed in [4], the percentage of time spent by the speaker looking at the face of the auditor increases steadily as the speaking turn approaches finality. [1] and [5] further elaborated on the role of facial expression and body placement in conversation and concluded that a number of turn-taking signals rely on participant location and facial cues. So, an interesting question to ask is what happens in terms of turn negotiation when people engaged in a conversation do not have access to gestures or facial cues. Assuming that prosody plays an important role in the organization of conversation and that it may be the only (or at least the most important) cue in telephone speech, would it be reasonable to expect that speakers in telephone conversation make a heavier use of it in order to compensate for the absence of non-linguistic cues? Would this variation be relevant from a perceptual viewpoint?

Earlier studies undertaken have aimed to isolate prosodic properties so that listeners were forced to decide whether a given string had reached a turn boundary without appealing to any physical behaviors or facial cues. Filtered speech is one way to test whether judges can distinguish turn-taking from turn-maintaining based on prosodic cues alone. Using this method, [6] compared the results of listening tests of face-to-face and non-face-to-face interaction, where judges decided whether a boundary was turn-beginning or turn-ending, based only on prosodic properties. Findings concluded that listeners vary with respect to how much they use intonation as a turn-taking signal.

More recent studies have demonstrated increased interest in the interaction of prosody and turn-taking [7]. Similar inquiries have incorporated syntactic properties of conversational units with prosodic elements of discourse [8] and [9]. [10] examines pitch peaks at syntactic boundaries as indicators of turn completions. [11] examines the pitch differences between turn ends and turn beginnings in three different settings: news broadcasts, magazine-style reporting and dialog.

In this paper, acoustic properties of Brazilian Portuguese turn boundaries are analyzed in both face-to-face and non-face-to-face interaction situations. One of the aims of the study is to determine whether speakers engaged in telephone conversation make a stronger use of prosody in order to compensate for the absence of non-linguistic cues. Perceptual tests are conducted in order to verify whether such variation is relevant from a perceptual viewpoint. If melodic cues are stronger in telephone speech, we expect this to have an effect on listeners’ prediction of turn-taking mechanisms.

2. Methodology

2.1. Material
The material used for this experiment was extracted from four spontaneous, natural telephone conversations and three face-to-face dialogues recorded for Project NURC [12]. In all cases but one, the dialogues were between two women. A total of 100 fragments from these interactions was first selected and included only female speech.

All fragments presented syntactic completion and were semantically neutral (i.e., they did not imply the motivation for a possible turn exchange, as, for example, a direct question would). These fragments were given to five experts in Brazilian Portuguese prosody, who had access to both the transcriptions and the digital audio files of all the excerpts. The experts were instructed to divide the fragments into intonation units and to indicate the type of boundary tone (low or non-low) at the end of the last intonation unit in the excerpts. In order for a boundary to be considered as “low” or “non-low” in the present work, three out of the five experts had to agree in their judgment. Most boundary tones were classified as either “low” or “non-low” unanimously. Only the last intonation unit of the excerpts was taken into account for both the acoustical and the perceptual analyses.

A total of forty representative excerpts, ranging from three to 19 seconds in length, were selected for the study. They are classified according to the types presented in Table 1.

2.2. Acoustical analysis
The speech files were digitized at a rate of 44,100 KHz with 16-bit resolution using Sound Studio (Felt Tip Software) speech-
Table 1: Description of the excerpts used in the study.

<table>
<thead>
<tr>
<th>channel</th>
<th>turn type</th>
<th>boundary tone</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>telephone</td>
<td>exchange</td>
<td>low</td>
<td>5</td>
</tr>
<tr>
<td>telephone</td>
<td>exchange</td>
<td>non-low</td>
<td>5</td>
</tr>
<tr>
<td>telephone</td>
<td>hold</td>
<td>low</td>
<td>5</td>
</tr>
<tr>
<td>telephone</td>
<td>hold</td>
<td>non-low</td>
<td>5</td>
</tr>
<tr>
<td>face-to-face</td>
<td>exchange</td>
<td>low</td>
<td>5</td>
</tr>
<tr>
<td>face-to-face</td>
<td>exchange</td>
<td>non-low</td>
<td>5</td>
</tr>
<tr>
<td>face-to-face</td>
<td>hold</td>
<td>low</td>
<td>5</td>
</tr>
<tr>
<td>face-to-face</td>
<td>hold</td>
<td>non-low</td>
<td>5</td>
</tr>
</tbody>
</table>

We also investigated whether boundary tone could have any effect on listeners’ decision. The results in Figure 5 show a trend to the expected direction. Figure 4 shows that prediction of turn change in telephone speech is much higher than in face-to-face speech. When considered separately, filtered speech had approximately the same amount of right predictions (hits) as non-filtered speech, as can be seen in Figure 3. The amount of wrong predictions (misses) was however greater for the filtered speech, suggesting that segmental information may have influenced listeners’ decision.

Since the result from filtered speech showed inconsistency, with a high number of wrong responses, we carried an analysis using only the non-filtered stimuli to see whether that would make a difference in terms of the general result. Although no statistically significant difference was found here as well, the numbers now show a trend to the expected direction. Figure 4 shows that prediction of turn change in telephone speech is much higher than in face-to-face speech.

We also investigated whether boundary tone could have any effect on listeners’ decision. The results in Figure 5 show a significant effect ($t(19) = -4.33, p < 0.001$) for boundary tone and listeners’ perception of turn transition in non-filtered speech (“low” for “hold”, “non-low” for ‘change’). On the other hand,
boundary tone is not a reliable cue for turn transition in filtered speech. A trend suggesting the use of low boundary tone for ‘hold’ and non-low boundary tones for ‘change’ does exist in here as well nonetheless. Although boundary tone alone may influence listeners’ perception of turn transition to some degree, no statistically significant correlation corroborated this assumption. Apparently, thus, listeners rely in other (non-melodic cues) for deciding whether a speaker will hold or give the conversational turn.

If only non-filtered speech ending in a low tone (conditions which previously showed some effect) is considered, the results are significant at least in telephone speech, as shown in Figure 6 (Wilcoxon signed rank, P < 0.02). The question to be asked thus is whether there are any other (prosodic) cues for the identification of turn management at play in telephone speech.

Figure 7 illustrates the correlation between listeners’ perception of turn change and the actual values of pitch range in the stimuli showing that higher values of pitch are associated with higher perception of turn change (216 Hz vs 157 Hz). This difference is not statistically significant though.

4. Conclusions

It is quite clear that the observable interaction between the prosodic cues under investigation and listeners’ judgements on turn management in conversation is not a straightforward one. The results in the present investigation corroborate what [6] and [18] found for English and Dutch, respectively, i.e., that melodic information by itself is for the most part insufficient to indicate turn transition to conversationalists.

No statistical basis was provided here for claiming that telephone conversationalists make use of different melodic cues to turn management. Listeners however seem to be able to predict turn management in telephone speech when some conditions are met, which would confirm - at least partially - the hypothesis that prosody plays a more important perceptual role in telephone speech.

Considering the results broadly - and the hypothesis that melodic cues alone would be sufficient for the perception of turn-taking mechanisms, the low agreements from subjects may be interpreted by taking into account the fact that those prosodic cues which were presented in the isolated utterances simply were not strong enough to signal either turn change or turn continuation for most of the utterances and most of the subjects. This problem relates to a larger issue, namely that of the importance of analyzing speech within the total context in which it is uttered. Researchers such as [19] and [20] have stressed the necessity of studying conversational phenomena within their actual contexts in order to have available all information which could possibly be affecting those conversationalists producing the phenomena. Isolating an object of study from its context may distort the analysis, since some contributing factors may be excluded from consideration in this way. One obvious candidate for such a neglected cue in the current test is inter-utterance pause, which has been stressed as an important organizing factor in conversation [21].

It is also important to note here the overriding optionality that exists at every level of conversation organization. Intonational cues may be used to manage turns in conversation, but not always they will be interpreted in the same way, especially when all other information in the surrounding conversational context
is taken into account. Therefore, it is desirable to test the role of intonation in turn management in broader contexts in order to fully understand how it actually functions in ongoing natural conversations.

5. References


