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 gestures. 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.*

channel	turn type	boundary tone	total
telephone	exchange	low	5
telephone	exchange	non-low	5
telephone	hold	low	5
telephone	hold	non-low	5
face-to-face	exchange	low	5
face-to-face	exchange	non-low	5
face-to-face	hold	low	5
face-to-face	hold	non-low	5

editing software. The data was subsequently analyzed under the speech-editing program Praat [13]. Pitch values in the signals were extracted automatically using the default fundamental frequency extraction algorithm in the program. The original pitch contours were then stylized by hand, in a semi-automatic process making use of both visual and auditory cues, thereby avoiding the interference of octave jumps and enabling us to smooth the contours ([14], [15] and [16]). The pitch range of intonation units was calculated by subtracting the value of minimum fundamental frequency from the value of maximum fundamental frequency, as measured at the nucleus of a vowel where both values occur in the same intonational unit.

2.3. Perceptual experiment

From the 40 original fragments, two of each category as described in Table 1 were filtered, totalling 16 filtered fragments. The filtering was done with Praat, through the synthesis of the intonation pitch contours as extracted with the program's native algorithm. The result was a sequence of pulses emulating the original pitch contour.

A total of 40 people (30 women, 10 men) participated in the perceptual experiment. All of them were native speakers of Brazilian Portuguese, mostly undergraduate students. None reported hearing difficulties.

Participants were given a brief introduction to the experiment, with an explanation of their role in it. They were instructed to listen to a given fragment, repeated continuously, and to predict whether there would be a turn change following it, to which they would answer YES.

No information regarding the intention of speakers was given to the participants. They were presented with an example of the filtered speech and its original counterpart and were finally informed that once their decision was made, the next fragment would be played. After five practice examples, the 40 randomized stimuli were presented.

3. Results

3.1. Acoustical analysis

Figure 1 suggests that, in general, pitch range tends to be a little higher when interactants in a conversation hold the turn. When interaction channel is considered separately, an interesting scenario is obtained. Interactants in telephone speech seem to use a different strategy to demonstrate their moves in conversation when it comes to pitch range: higher values are often associated with turn changing transitions. Although Figure 1 presents a trend in these opposing directions, no statistical significant effect could be found to validate it.

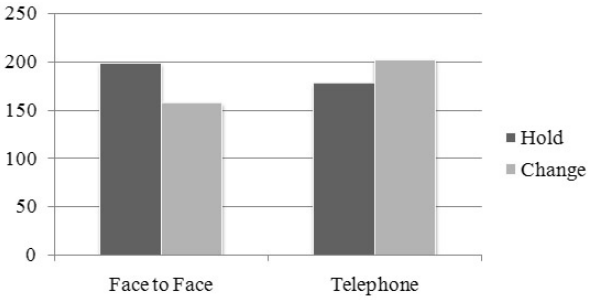

Figure 1: *Pitch range values (in hertz) of transition types (hold or change) per channel (face-to-face or telephone).*

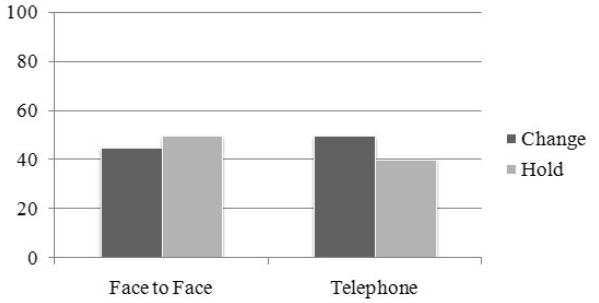

Figure 2: *Subjects' prediction of turn-taking mechanisms in face-to-face and telephone speech (in percentage).*

3.2. Perception tests

In order to measure listeners' agreement with respect to turn transition, kappa coefficient was used [17]. In the present study, the coefficient was found to be lower than 0,6, which indicates that listeners do not agree with each other when it comes to judging the transitional state of an utterance in a conversation.

Results reported in Figure 2 indicate that listeners were not able to consistently determine whether a given utterance is followed by a turn hold or a turn change. Statistical analyses showed no interaction between listeners' judgement and right answers. In telephone speech, only nine fragments (out of twenty) received the right answers in an above than average scale. In face-to-face speech, this number increased to ten.

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,

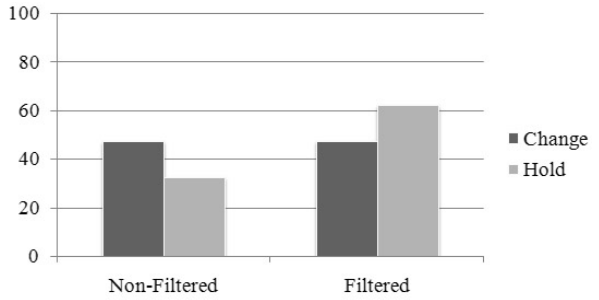

Figure 3: *Subjects' prediction of turn-taking mechanisms in filtered and non-filtered stimuli (in percentage).*

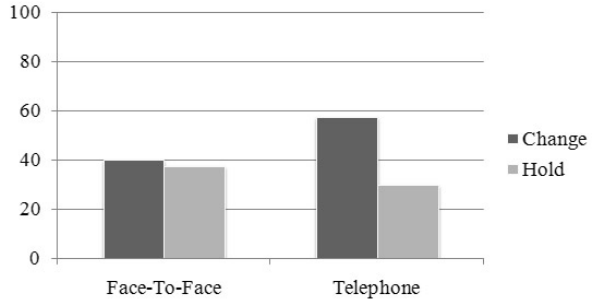

Figure 4: *Subjects' prediction of turn-taking mechanisms in telephone and face-to-face speech (non-filtered stimuli, in percentage).*

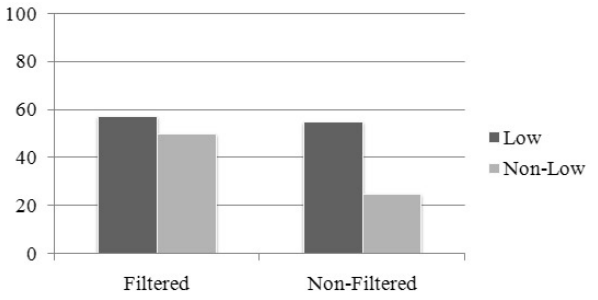

Figure 5: *Subjects' prediction of turn change with fragments ending in low or non-low boundary tone in filtered and non filtered speech (in percentage).*

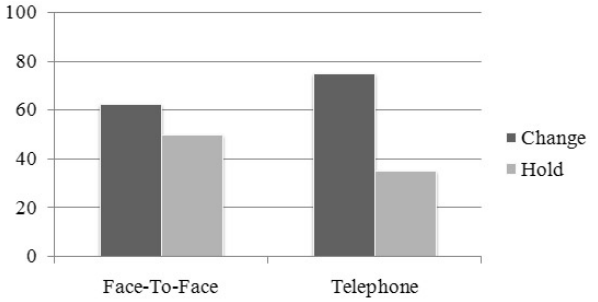

Figure 6: *Subjects' prediction of turn change in telephone and face-to-face speech (non-filtered stimuli ending in low tone, in percentage).*

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

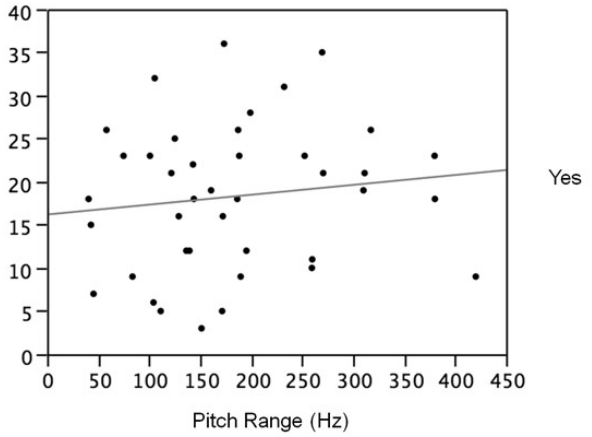

Figure 7: *Correlation of listeners' perception of turn change and pitch range.*

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

[1] Sacks, H.; Schegloff, E.; Jefferson, G., 1974. A simplest systematics for the organization of turn-taking in conversation. *Language*, 50, 696-735.

[2] Duncan, S. Jr., 1972. Some signals and rules for taking speaking turns in conversations. *Journal of Personality and Social Psychology*, 23(2), 283-292.

[3] Walker, G., 2004. *The phonetic design of turn endings, beginnings, and continuations in conversation.* PhD thesis, University of York.

[4] Wiemann, J.; Knapp, M., 1975. Turn-Taking in conversations. *Journal of Communication*, 25, 75-92.

[5] Goodwin, C., 1981. *Conversational organization: interaction between hearers and speakers.* New York: Academic Press.

[6] Schaffer, D., 1983. The role of intonation as a cue to turn taking in conversation. *Journal of Phonetics*, 11, 243-257.

[7] Hidalgo-Navarro, A., 1999. Procedures de demarcation dans l'alternance des prises de parole: Interaction, syntaxe et prosodie. *Faits de Langues: Revue de Linguistique*, 13, 110-124.

[8] Auer, P., 1996. On the prosody and syntax of turn-continuations. In *Prosody in Conversation.*, E. Couper-Kuhlen, M. Selting (eds.). Cambridge: Cambridge University Press, 57-100.

[9] Selting, M., 1996. On the interplay of syntax and prosody in the constitution of turn-constructional units and turns in conversation pragmatics. *Quarterly Publication of the International Pragmatics Association*, 6(3), 371-388.

[10] Schegloff, E., 1998. Reflections on studying prosody in talk-in-interaction. *Language and Speech*, 41(3-4), 235-263.

[11] Mindt, I., 1999. Prosodic cues at speaker turns. *Papers from the Twentieth International Conference on English Language Research on Computerized Corpora (ICAME 20), Freiburg im Breisgau.* Rodopi: Amsterdam.

[12] Sá, P.; Cunha, D.; Lima, A.; Oliveira, M. (eds.), 1997. *A linguagem falada culta na cidade do Recife: diálogo entre informante e documentador.* Editora Universitária: Recife.

[13] Boersma, P., Weenink, D., 2003. Praat: doing phonetics by computer (Version 4.1.5). http://www.praat.org/

[14] Nooteboom, S., Kruyt, J., 1986. Accents, focus distribution, and the perceived distribution of given and new information: an experiment. *Journal of the Acoustical Society of America*, 82(5), 1512-1524.

[15] Sluijter, A., Terken, J., 1993. Beyond sentence prosody: paragraph intonation in Dutch. *Phonetica*, 50, 180-188.

[16] van Donzel, M., 1999. *Prosodic aspects of information structure in discourse.* PhD Thesis, University of Amsterdam.

[17] Carletta, J., 1996. Assessing agreement on classification tasks: the kappa statistics. *Computational Linguistics*, 22(2), 249-254.

[18] Caspers, J., 1998. Who's next? The melodic marking of question vs continuation in Dutch. *Language and Speech*, 41, 375-398.

[19] Gunter, R., 1972. Intonation and relevance. In *Intonation: Selected Readings.* D. Bolinger (ed.). Harmondsworth: Penguin: 194-215.

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

[21] Erickson, F., 1982. Money tree, lasagna bush, salt and pepper: social construction of topical cohesion in a conversation among Italian-Americans. In *Analyzing Discourse: Text and Talk*, D. Tannen (ed.). Washington: Georgetown University Press: 43-70.