Acoustic Cues for Classifying Communicative Intentions in Dialogue Systems

Michelina Savino & Mario Refice

D.E.E. – Politecnico di Bari, ITALY
esavino@poliba.it, refice@poliba.it

Abstract. Filled pauses are normally used as a planning strategy: they signal speaker's intention to hold the floor in a conversation. They are normally realised by inserting a vowel (optionally followed by a nasal), but in Italian they can be produced by lengthening the final vowel of a word. Word final lengthening filled pauses are then an intermediate category between lexical and non-lexical speech event. In human machine interaction, the system should be able to discriminate between a "default" lexical speech event and one characterised by a word final lengthening for planning strategy: in this second case, the related communicative intention has to be additionally recognised. Our preliminary investigation shows that duration and F0 shape are reliable acoustic cues for identifying word final lengthening filled pauses in a variety of Italian.

1 Introduction

One of the main challenge in the development of dialogue systems is the attempt to "naturalising" interaction as much as possible, i.e. allowing stronger control of dialogue modality by the human being and not by the system (as it is presently the case). For example, in turn-taking the system should be able to "recognise" speech events segnalising human communication intentions of "giving the floor" from those which are simply pauses within a turn. Most of these phenomena are not reliably predictable, and therefore not easily managed if the system lacks the needed knowledge. Most of the present dialogue systems make use of pauses (i.e. no speech production by the user) as a "signal" for switching the floor back to the system itself. This simple technique results, in several cases, in the overlapping of human and machine speech, with the consequent loss of control of the planned interaction process. On the other hand, filled pauses are not "interpreted" as user's intention of "holding the floor", but simply as lexical speech items in themselves.

The purpose of this paper is to discuss possible solutions for coping with the mentioned problem, trying to discriminate between different uses of pauses. In particular, preliminary results on attempt of modelling (a special type of) filled pauses will be given, basing on data collected and properly annotated in a spontaneous, dialogue-based speech database of Italian regional varieties under development within a national project.

2 Communicative and Non-Communicative Speech Events

In human verbal communication, it is possible to identify two main typologies of speech events: lexical speech events and non-lexical speech events. The latter can be further classified in:

1. Non-lexical speech events conveying communicative intentions like feedback and turn-taking, such as filled pauses;
2. Non-lexical speech events which do not convey any communicative intentionality, such as coughing, sneezing, etc.

Despite this clearly different function, most of the traditional work (expecially in ASR and man-machine interaction in general) has considered both classes as one category, that of "disfluency phenomena", sharing the same function of class 2, and, as such, of less interest in terms of modelling with respect to that of lexical speech events. Only recently, attention has been paid to the description, classification and formalisation of the above mentioned non-lexical communicative speech events [1], mainly with reference to ASR performance improvement [2], [3]. In this work we try to make a step further in modelling filled pauses, by recovering additional information, in terms of communicative intentionality conveyed by this kind of disfluency, in dialogue system applications.

3. "Word Final Lengthening" Filled Pauses

Filled pauses are normally used as a planning strategy: they signal speaker's intention of holding the floor in a conversation. From the segmental/acoustic point of view, they are realised by inserting a central vowel, optionally followed by a nasal (less common), but their segmental realisations may be language-specific. For example, in Italian - where almost all words end by a vowel - speaker do not need to insert a "spurious" vowel-like segment in their speech: they can achieve the same by simply lengthening a word ending vowel. For the same fonotactic reasons, if the word ends by a consonant, then a central vowel schwa is added and prolonged. "Word final lengthening" filled pauses are then a sort of intermediate category, half way between lexical and non-lexical speech events, since they consist of a lexical event being "affected" by a non-lexical speech phenomenon. In human-machine interaction for Italian, then, the system should be able to distinguish between a "default" lexical speech event and a lexical speech event characterised by a "word final lengthening" filled pause. In the latter case, a particular communicative intention (i.e. holding the floor) has to be additionally recognised by the system.

At a first glance, duration seems to be the main acoustic cue to be used in the above mentioned classification task. Yet duration reliability cannot be taken for granted without detailed investigations, since in determining threshold values one has to consider other possible cases of word final lengthening, i.e. those occurring at (major) prosodic boundaries, and conveying intentionality of different types. With this respect, we hypothesised also that an additional acoustic cue which might be relevant for discriminating word final lengthening filled pause is F0 shape.

4 “Word Final Lengthening” at Prosodic Boundaries

Word final lengthening is also possible, with different degrees of prolongation, at (major) prosodic boundaries. The higher degrees of lengthening are normally associated to cases of phrase final nuclear stressed syllables, i.e. when a complex tonal sequence (pitch accent + boundary tones) has to be realised on one syllable (tonal crowding). Theoretically, it may be postulated that the more complex the movement to be realised, the stronger the degree of lengthening to be found. Actually, languages may adopt either a compression strategy (i.e. realisation of the whole complex tonal sequence on one syllable with its consequent prolongation) or a partial/total truncation one (i.e. partial/total truncation of the sequence: in this case the movement is not completely realised) [4]. In Italian, prevailing of one or the other strategy is variety-depending. Bari Italian speakers - being the regional variety under investigation that of Bari - may use both of them, depending on speaking style, speaking rate, or some added paralinguistic meaning [5]. In analysing our Bari Italian speech material, then, we expect to find cases of phrase final stressed syllable with (possibly) different degrees of prolongation. In this preliminary analysis, we considered them as belonging to one broad category.

5 Method

In order to discriminate betwen “default” lexical speech events and those characterised by the lengthening of the word ending vowel, a number of duration measurements has been carried out. In our analysis, default or unmarked cases of vowel duration at the end of a word are those in stressed open syllables which are not phrase final. Such default durations have been compared with those in two marked contexts: a) (degrees of) prolongation at prosodic boundary, and b) prolongation for planning strategies. Therefore, duration measurements have been performed on the following types:

1. vowels in word (but not phrase) final stressed open syllables (“default”);
2. vowels in phrase final stressed open syllables;
3. vowels in word final stressed and unstressed open syllables with “planning lengthening”.

Moreover, F0 shapes of the three above mentioned typologies have been described and compared. In our analysis, both monosyllabic and polisyllabic words have been taken into account.

6 Speech Material

Analysed speech material consists of three spontaneous Map Task [6] dialogues between pairs of Bari Italian speakers, which is a subset of a corpus of some regional varieties of Italian (namely those of Bari, Naples and Pisa) under development within the national project AVIP (Archivio di Varietà dell’Italiano Parlato, Spoken Italian

Varieties Archive). The speech material has been ortographically transcribed, segmented and annotated (both at segmental and suprasegmental levels), including non-lexical speech events. A special label was used for coding cases of "word final lengthening for planning purposes".

7 Preliminary Results

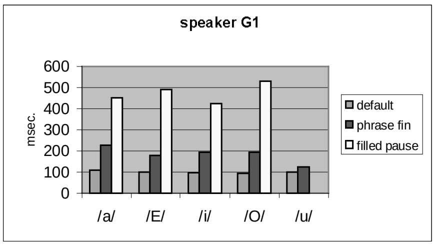
7.1 Duration

Duration mean values of all five vowel types (in stressed open syllable) in the three above mentioned contexts for the six speakers are shown in Figures 1, 2, 3, 4, 5, 6, respectively. Letters G and F identify speaker role within each dialogue (instruction Giver or instruction Follower), and number refer to the pair of maps used (each pair of map having different path and different landmarks names).

Preliminary results, within the statistical significance of the available data, confirm that mean duration of phrase final stressed vowels are sistematically higher than those in "default" position for all speakers. Some inter- and intra-speaker variability within the category "final lengthening at prosodic boundaries" is to be noted, which can be related to the different degrees of prolongation, as discussed in section 4. Since we are interested, at this stage, in determining a threshold value between the two broad categories "word final lengthening at prosodic boundaries" and "word final lengthening as filled pauses", a further sub-classification within such variability has not been considered in the present work.

Our results show also that mean duration values of word ending vowels with "planning lengthening" are sistematically higher than those in the two remaining categories, for all vowel types[1]. As vowel /u/ is concerned, no cases of final lengthening caused by planning strategies was found in our dialogues. This is not surprising, since /u/ ending words are less common in Italian. On the basis of these results, we can give some indications of duration ratios among the three categories considered. Figure 7 shows ratio, for each vowel type, between mean duration of category "word final lengthening for planning strategies" and the default category "word (but not phrase) final", for all speakers: vowels in word final stressed and unstressed open syllables can be classified as a "planning lengthening" phenomenon if they are at least three times longer than in "default" cases. Figure 8 shows ratio, for each vowel types, between mean duration of category "word final lengthening for planning strategies" and "word final lengthening at prosodic (major) boundaries", for all speakers: in this last case, vowel with planning lengthening are at least two times longer than those at prosodic boundaries.

[1] Stressed and unstressed vowels mean durations for this category were first computed separately; since they did not show any statistically significant difference, both types have been merged into one group.



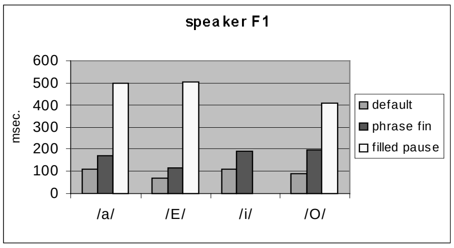
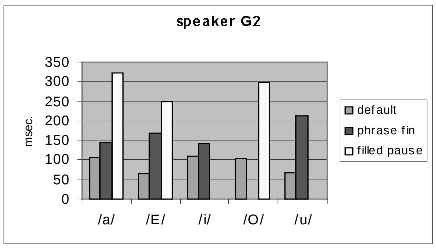


Fig. 1. and **Fig. 2.** Mean durations, for all vowel types (when found), in word final ("default"), at prosodic boundary ("phrase fin") and as filled pause positions for Giver (G1) and Follower (F1) in Map Task dialogue n.1



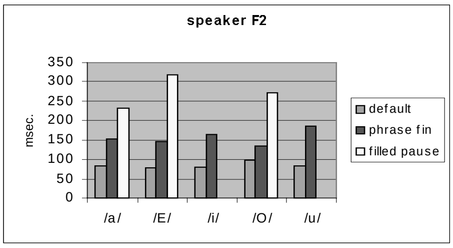
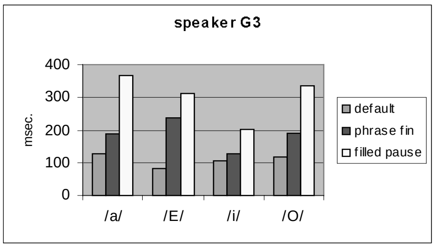


Fig. 3. and **Fig. 4.** Mean durations, for all vowel types (when found), in word final ("default"), at prosodic boundary ("phrase fin") and as filled pause positions for Giver (G2) and Follower (F2) in Map Task dialogue n.2



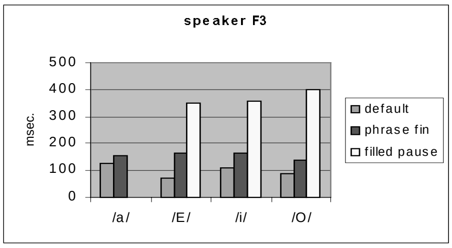


Fig. 5. and **Fig. 6.** Mean durations, for all vowel types (when found), in word final ("default"), at prosodic boundary ("phrase fin") and as filled pause positions for Giver (G3) and Follower (F3) in Map Task dialogue n.3

Table 1. % of F0 shape types in filled pauses and phrase final vowels for all speakers (R=Rise, F=Fall)

	R,F	RF, FR, RFR	LEVEL
phrase-final	76%	22%	2%
filled pauses	-	-	100%

7.2 F0 Shape

We considered also F0 shape as a possible additional acoustic cue in classifying the above mentioned types of filled pauses, expecially with respect to that of lengthening at prosodic boundaries. Results of our F0 shape analysis for both the above mentioned categories are shown in Table 1. It can be noticed that a level melodic contour (i.e. continuity of F0 values throughout vowel duration) strongly characterises vowels lengthened for planning strategies, in contrast with the variability (also in terms of complexity of the shape) of the "prosodic boundary lengthening" category, where this variability is related to a number of different communicative functions.

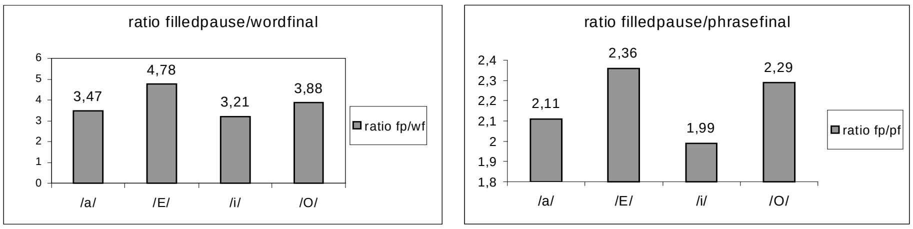


Fig. 7. and **Fig. 8.** Ratio of filled pauses on word-final (default) vowels mean duration values (left) and ratio of filled pauses on phrase-final vowels mean duration values (right)

8 Conclusion and Future Work

Our preliminary results suggest that - by means of simple decision rules in a man-machine dialogue system - duration and F0 shape can be used as reliable acoustic parameters in classifying communicative intentions at least in some broad classes. More specifically, these cues can help in detecting and interpreting typical Italian word final lengthening filled pauses as the intention of "holding the floor" by the speaker. A further sub-classification of communicative intentionality, within the broad category "final lengthening at prosodic boundaries" will be carried out as soon as a larger corpus will be available.

References

1. Gybbon D. & Shu-Chuan Tseng: Toward a formal characterisation of disfluency processing, in: ICPhS99 sat. meet "Disfluency in Spontaneous Speech" Proc. (1999) 35-38
2. O'Shaughnessy D.: Better detection of hesitations in spontaneous speech, in: ICPhS99 satellite meeting "Disfluency in Spontaneous Speech" Proc. (1999) 39-42
3. Pakhomov S. & Savova G.: Filled pause distribution and modelling in quasi-spontaneous speech, in: ICPhS99 sat. meeting "Disfluency in Spontaneous Speech" Proc. (1999) 31-34
4. Ladd R.D.: Intonational phonology, Cambridge Univ. Press, Cambridge, (1996)
5. Refice M, Savino M, Grice M.: A contribution to the estimation of naturalness in the intonation of Italian spontaneous speech, in: Eurospeech 97 Proc. (1997) 783-786
6. Brown G., Anderson A., Yule G., Shillcock R.: Teaching talk, Cambridge Univ. Press, Cambridge (1983)