ProsoReportDialog: a tool for temporal variables description in dialogues

Jean-Philippe Goldman

Linguistic Department, University of Geneva, Switzerland

jeanphilippegoldman@gmail.com

Abstract

Temporal variables were initiated by Grosjean in 1972 [1] who defined in details several dimensions in timing and rhythm in order to measure and compare these characteristics in different languages or various speaking context. Aside this approach dedicated to monologues, some studies applied notions from conversational domain (as in Sachs et al. 1974 [2]) to dialogues corpus in an automatic way.

Our contribution extends this work in both directions. (i) We suggest gathering these two approaches in an automatic tool for temporal variables description in dialogues. (ii) We compare these variables in a corpus of 35 dialogues to show their differences according their situational features.

Index Terms: tool, prosody, dialog, tools, resources, analysis, and speech prosody

1. Introduction

Temporal variables in speech have been studied extensively, in the domain of speech synthesis to improve the naturalness of the synthetic speech [3], in man-machine dialogue modelling [4], or in the field of conversation analysis, using a qualitative [5] or quantitative [6] approach within corpus linguistics. Finally, a number of studies have focused on temporal aspects in a descriptive approach [7][8][9], or in a contrastive approach [10] or even to study phenomena as hesitations [11].

These numerous studies deal with a wide range of languages and speaking situations, providing measures that are not often comparable because of the procedures used to extract and analyse them. To mention only few debated points: 1. Should the so-called "micro-pauses" (pauses smaller than a threshold) be considered or not (see [8] for a review of this bias) 2. Should the pause duration be log-transformed or not (see [4]). 3. Under which threshold of time difference should two turn boundaries (start or end point) be considered as simultaneous ?

After recalling some basics of temporal variables, we present an automatic tool to measure the temporal variables and apply it to a corpus of 35 various dialogs. We limit our current work to dialog (i.e. with exactly 2 speakers) leaving a third, a forth speaker for a later study.

2. Temporal variables

2.1. In monologues

More than forty years after its publication, the pioneering work of [1] remains a reference for defining the temporal variables of the oral language. The total (or speech) time is composed of the articulation time (or phonation) and pause time, from which are derived the articulation and pause ratios (as a percentage of speech time) as well as the articulation rate (in syllables per second). Some other notions are taking into account like the number and the length of the speech sequences separated by pauses.

On the top of this, additional secondary variables are also considered such as filled pauses (hesitation and syllable lengthening), repetitions and false starts. These variables require manual annotation and thus are often ignored in studies on large corpora.

2.2. In dialogues

The analysis of conversations between two or more speakers makes the study of temporal variables more difficult; especially if overlapping speech occur. The notion of speech turn, which seems to be central, is extremely difficult to implement in an automatic analysis system, as this unit is the result of a dynamic analysis of how the speakers combine turn constructional units (TCU's) incrementally to produce what will be considered in context, as a turn of speech [12][13].

For automatic annotation, the notion of verbal production (VP) as a sequence of syllables assigned to a unique speaker should be preferred to speech turn ([14]). Verbal production can be long sequence syllables but in some cases a brief backchannel output, occurring within a pause or overlapping with the other speaker.

The silent pauses may occur within the VPs of a speaker (so-called within- or intra-speaker pause) or between the end of the verbal output of a speaker and the beginning of the next speaker (between- or inter-speaker pause or gaps). The former can simply be called "pauses" if the latter are referred as "gaps".

The transition from a speaker to another can occur without any gap or overlap (the famous no-gap-no-overlap as in [2]), but often leads to a speech overlap of speech which, at most times, is not perceived as an interruption of the speaker being but as a slightly early transition [15].

The most complete list of turn change patterns is provided by [16], identifying 10 cases. This model has been often simplified (see [4][6]) because its implementation requires manual annotation of some phenomena, such as backchannels.

3. Procedure

To derive the temporal variables of a dialog, the main relevant information can be embedded in a single tier with speaker annotation, showing which speaker is speaking at every moment. Pauses, gaps and overlapping segment are also indicated.

The same information can possibly lie within several tiers (one tier per speaker). In this implementation, multiple tiers are firstly merged into a unique one.

As mentioned before, silent pauses are split into intra-speaker pause and inter-speaker gaps. If a silence is surrounded by two VPs of the same speaker, it is a pause. If a speaker transition occurs, then it is a gap. This pause-gap difference becomes a problem when an overlapping interval is adjacent to a pause. This happens when the two speakers start or stop speaking simultaneously.

[17] makes a systematic distinction, suggesting the *Instigator* and *Owner* status for each pause: " the *instigator* of a silence is the speaker who last spoke before the silence occurred (or who last spoke alone, in cases of a simultaneous end of speech); the *owner* of the silence is the speaker who breaks the silence (or the instigator, in cases of simultaneous start of speech); a *gap* is a silence with a different instigator and owner (aka *inter-speaker silence*); and a pause is a silence with the same instigator and owner (aka *intra-speaker silence*) "

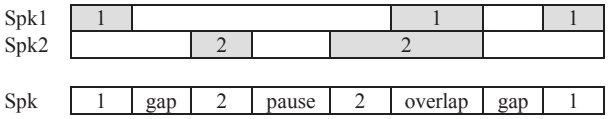

Figure 1. Example for separate tiers (above) and merge tier (below)

On the basis of a merged speaker tier where verbal productions, breaks, gaps and overlaps are distinguished, the tool offers the following measures, in order to depict as simply as possible the composition of the dialog:

- Recording Time
- Speech time (excluding side-breaks)
  - Articulation time
    - Exclusive articulation
    - Overlap (initiated, with/without transition)
    - Cumulated articulation
  - Silence time
    - pauses (intra)
    - gaps (inter)

These dimensions are shown with their duration (in seconds) and their count or frequency (number of pauses, number of verbal productions, number of overlap sequences). In addition, complex variables are derived as:

- ratios (as a percentage of the speech time)
- mean durations
- rates (per second or per minute)

For instance, the articulation ratio is identical to the so-called "rapport TA-TL" (so-called "rapport Temps Articulation-Temps de Locution") in [1]. These measures are also detailed for each speaker. It should also be mentioned that the notion of cumulated articulation can be greater than 100% of the speech time. In other words, the exact articulation time is added for each speaker (i.e. overlap segments count more than once).

Moreover, the speaker initiating an overlapping interval (e.g. starting of VP while the other is currently speaking) can be clearly identified. Besides, this overlap segment leads to a turn change or not, it is counted with or without transition. In the second case (no transition), the overlap could be a backchannel VP or an aborted try of turn taking. The distinction of these latter cannot be done automatically at this moment.

In practice, the tool takes a TextGrid file with a speaker tier as input, and offers four different outputs. In the first two, all these above measures are displayed in a shortened version within the Info window of Praat (which can be saved in a text file). It is a simple overview of some measures. One is articulation-oriented as the other is speaker-oriented as can be seen below:

```
######Processing TextGrid foot0...
Speech time            299
  -articulation        214            (71.6%)
    -overlap             11           (3.6%)
    -exclusive artic. 203             (67.9%)
      -spk 1               79         (26.4%)
      -spk 2               124 (41.5%)
  -silence             85             (28.4%)
    -gap                 24           (8.0%)
    -pause               61           (20.4%)
      -spk 1               36         (12.0%)
      -spk 2               25         (8.4%)


######Processing TextGrid foot0...
Speech time            299
  -spk 1               126            (42.1%)
    -exclusive artic.    79           (26.4%)
    -overlap             11           (3.7%)
    -pause               36           (12.0%)
  -spk 2               160            (53.5%)
    -exclusive artic.    124          (41.5%)
    -overlap             11           (3.7%)
    -pause               25           (8.4%)
  -gap                 24             (8.0%)
```

The third output is an extended version as below:

```
######Processing TextGrid foot0...
Tier speaker found in TextGrid : 4
Recording time          300.012
Speech time             299.134 (side.pauses
excl)
Silence time            85.042  (28.4%/speechime)
Pause (intra) time      61.402  (20.5% )
Gap (inter) time        23.640  (7.9 %)
Articulation time       214.092 (71.6%)
Cumulated articulation 224.725 (75.1%)
Overlap time            10.633  (3.6%/speech ime)
                                (5.0%/art.time)
Exclusive articulation 203.459 (68%/speech time)
                                (95%/art.time)
Nb of pauses            84 (dur:1.012;rate:16.8)
Nb of intra pauses      50 (dur:1.228;rate:10.0)
Nb of pauses inter      34 (dur:0.695;rate:6.8)
Nb of VP                120(dur:1.784;rate:24.1)
                           (min:0.05;max:6.238)
Nb of overlaps          25 (dur:0.425;rate:5.0)
Nb initiated overlap    28 (w/o transition 13)


###Speaker #1###
Articulation time       90.023  (42.0%/art.time)
                                (40.1%/cumul.art)
Exclusive articulation 79.390  (26.5%/speech t.)
                                (37.1% /art.t.)
Pause (intra) time      36.386  (12.2%/speech t.)
Nb of VP                56 (dur:1.608;rate:24.1)
                           (min:0.05;max:5.813)
Nb overlaps    25          (rate: 5.0)
Nb initiated overlap    12 (w/o transition: 6)
Nb of intra pauses      21 (dur:1.733;rate:4.2)



###Speaker #2###
Articulation time       134.702 (62.9%/art.time)
                                (59.9%/cumul.art)
Exclusive articulation 124.069 (41.5%/speech t.)
                                (58.0%/art.t)
Pause (intra) time      25.016  (8.4% /speech.t)
Nb of VP                64 (dur:2.105;rate:24.1)
                           (min:0.05;max:6.238)
Nb of overlaps          25 (rate: 5.0)
Nb initiated overlap    16 (w/o transition:7)
Nb of intra pauses      29 (dur:0.863;rate:5.8)
###
```

Finally, the 4th option outputs all the measures of the full report in a table (tab-separated or csv) rather in a text window. In this case, many dialogs can be analysed at once and the results are represented in columns for all the speech files. The table format permits further analysis.

The tool has been developed as a Praat plugin and includes with some extra tools to manipulate intervals tiers such as:

- **Merging interval tiers:** if speaker tiers are separated, this tool produce a unique speaker interval tier
- **Report pauses from a syllabic tier to the speakers tier:** if a syllabic tier exists, pauses can be derived and added to the speaker tier.
- Merging similar consecutive intervals: if several pause intervals or several same-speaker intervals exist, they can be merge as one, avoiding future wrong measures.

4. Corpus

In this part, we apply the described methodology to a group of 35 extracts representing various speaking styles in different activities and situations. The total duration is 2 hours and 40 minutes. To cite a few examples, there are radio interviews, radio news dialogs as well as sports live report or map task dialogues.

We annotated each dialog according to a set of situational features [17][18] to allow further study and comparison. Our hypothesis is that these situational features may yield differences in the temporal variables. These features are the degree of interactivity (interactive, semi-interactive, non-interactive), the degree of preparation (spontaneous, semi-prepared, prepared) and the degree of media use as in the next Table.

	Interactive (total = 14)	semi-inter. (total =16)	non-inter. (total = 5)
Spontan. (total= 16)	D0009,D2008 **foot0,foot1, foot31,**intlib1, intlib2,intlib3	D0003,D0005 D1003,D2004	D0007 D0008 D0017 D0020
Semi-prepared (total= 16)	D0004,D0006 D2001 D2002 D2010 D2012	D1001,D1002 D2009,infor1 infor2,intfor3 intrad1,intrad2 intrad3,intrad4	
Prepared (total = 3)		D2005 D2006	**D2013**

*Table 1. Distribution of the 35 speech samples according the 3 situational features. The media feature is indicated as non-media (total = 16), secondary media (total = 15), **media** (total = 4)).*

5. Results

In the following table, some mean measures are represented for the 35 dialogs.

Variable	Mean	Std deviation
Speech time	275.0 (s)	110.3
Articulation ratio	79.2 (%)	7.9
Overlap ratio	4.2 (%)	4.7
VP duration	10.4 (s)	8.7
VP duration spk1	16.7 (s)	13.5
Speech ratio spk1	77.7 (%)	14.6
VP duration spk2	3.3 (s)	2.7
Speech ratio spk2	20.2 (%)	14.4
Gap duration	0.7 (s)	0.6
Gap ratio	5.4 (%)	3.3

Table 2. Vital statistics for the corpus of 35 dialogs

As a first result, we plotted the speech ratio of both speakers as in Figure 2. The 35 extracts are scattered along a diagonal as the sum of their speech ratio is supposed to be 100%. The items below the line have more gaps than overlaps.

The "interaction" feature is divided into three categories: non-interactive (red), semi-interactive (green) and interactive (blue). Each of these three categories gather in groups although interactive (blue) and non-interactive (red) ones seem to superimpose leaving aside the semi-interactive.

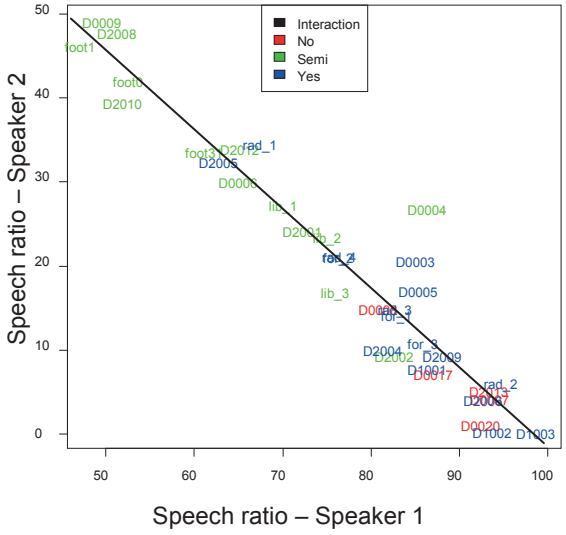

Figure 2. A corpus of 35 dialogues represented as speech ratio of speaker1 vs. speech ratio of speaker2

6. Discussion

This first attempt to automatically extract temporal variables in dialogs showed a high number of unexpected problems. Some questions still need further investigations. However, further developments are already in preparation like speech rate as well as an estimation of dynamic variation of the temporal variables along the total speech recording.

This tool is freely available and is distributed under this website:

http://latlntic.unige.ch/phonetique

7. Acknowledgements

This work is part of the Swiss-FNS project "Prosodic and linguistic characterization of speaking styles: semi-automatic approach and applications" (fund n°100012_134818).

8. References

[1] Grosjean, F. & A. Deschamps (1972). Analyse des variables temporelles du français spontané. Phonetica, 26, pp.129‑156.

[2] Sacks, H., E. A. Schegloff & G. Jefferson. (1974). A Simplest Systematics for the Organization of Turn-Taking for Conversation. Language, 50(4), pp. 696‑735. Smith, J. O. and Abel, J. S., "Bark and ERB Bilinear Transforms", IEEE Trans. Speech and Audio Proc., 7(6):697-708, 1999.

[3] Zellner, B., 1998. Caractérisation et prédiction du débit de parole en français Une étude de cas.

[4] Heldner, M., & Edlund, J. (2010). Pauses, gaps and overlaps in conversations. Journal of Phonetics, 38, 555-568. doi: 10.1016/j.wocn.2010.08.002

[5] Auer, P.,Couper-Kuhlen, E., & Müller, F. (1999). Language in time: The rhythm and tempo of spoken interaction. New York: Oxford University Press.

[6] Ten Bosch, L., N. Oostdijk & L. Boves (2005). On temporal aspects of turn taking in conversational dialogues. *Speech Communication*, 47:1–2, pp. 80‑86

[7] Duez, D. (1987). Contribution à l'étude de la structuration temporelle de la parole en français, PhD thesis Université de Provence.

[8] Campione, E. & Véronis, J. 2002. A large-scale multilingual study of silent pause duration. *SP-2002*, 199-202

[9] Goldman, J. et al., 2010. Prominence perception and accent detection in French: a corpus-based account. In Speech Prosody.2010, Chicago.

[10] Grosjean, F. & Deschamps, A., 1975. Analyse contrastive des variables temporelles de l'anglais et du français: vitesse de parole et variables composantes, phénomènes d'hésitation. Phonetica, 31, pp.144–184.Candea 2000

[11] Selting, M. (2005) Syntax and prosody as methods for the construction and identification of turn-constructional units in conversation. In: Hakulinen, Auli and Margret Selting (eds.), Syntax and Lexis in Conversation: Studies on the use of linguistic resources in talk-in-interaction . 2005. (pp. 17–44)

[12] Mondada, L. (2008). L'interprétation online par les co-participants de la structuration du tour in fieri en TCUs: évidences multimodales. Tranel 48, pp.7‑38.

[13] Groupe ICOR. (2006). Glossaire. Site CORINTE http://icar.univ-lyon2.fr/projets/corinte/[consulté le 7/5/2013]

[14] Jefferson, G. (1983). Notes on some orderliness of overlap onset. Tilburg Papers in Language and Literature 28, Department of Linguistics, Tilburg University.

[15] Weilhammer, K. & Rabold, S. (2003). Durational Aspects in Turn Taking, ICPhS, p. 931-934.

[16] Edlund, J., Heldner, M. & Hirschberg, J., 2009. Pause and gap length in face-to-face interaction. In *Proceedings of Interspeech 2009*. Citeseer, pp. 2779–2782.

[17] Koch, P., & Oesterreicher, W. 2001. Langage parlé et langage écrit. (G. Holtus, M. Metzeltin, & C.Schmitt, Éd.) Lexikon der romanistischen Linguistik (LRL). Tübingen: Niemeyer.

[18] Simon, A.C., A. Auchlin, M. Avanzi & J.-Ph. Goldman. (2009) Les phonostyles: une description prosodique des styles de parole en français. In: Abecassis, M. & G. Ledegen, Les voix des Français. En parlant, en écrivant, Berne: Peter Lang, 71-88.