Automatic Detection and Prediction of Topic Changes Through Automatic Detection of Register variations and Pause Duration

Céline de Looze, Stéphane Rauzy

Automatic Detection and Prediction of Topic Changes Through Automatic Detection of Register variations and Pause Duration

Céline De Looze & Stéphane Rauzy

Laboratoire Parole et Langage, CNRS et Université de Provence, Aix-en-Provence, France
{celine.delooze; stephane.rauzy}@lpl-aix.fr

Abstract

In this article a clustering algorithm, allowing the automatic detection of speakers' register changes, is presented. Together with automatic detection of pause duration, it has shown to be efficient for the automatic detection and prediction of topic changes. The need to take into account other parameters such as tempo and intensity, in the framework of Linear Discriminant Analysis, is proposed in order to improve the identification of the topic structure of discourse.

Index Terms: register variations, pause duration, topic changes, automatic detection and prediction.

1. Introduction

In this article, a clustering algorithm is presented, which allows the automatic detection of a speaker's changes of register. Developed to be implemented in the automatic coding of intonation models, it has been used, together with automatic detection of pause duration, to detect and predict topic changes automatically.

We assume that two types of fundamental frequency patterns need to be distinguished: on the one hand, local pitch characteristics corresponding to changes in the phonological representation of intonation and, on the other hand, more global pitch changes determined by variations in register as defined by key (or level) and span (or range). Assuming that a speaker's register may vary, especially when analyzing spontaneous speech, and that these variations may convey linguistic, extra-linguistic as well as para-linguistic functions, will certainly improve the (automatic) description of intonation patterns.

More particularly, register is reported to throw light on the informational organisation of discourse structure, the information weight carried by the discourse element compared to its preceding or following neighbour as well as the hierarchical dimension and relational organisation of linguistic units. Discourse units are more or less glued together depending on their semantic/pragmatic relations so as to form a coherent whole. It has been shown, for example, in many languages, that the first sentences of paragraphs in reading tasks are uttered with higher register than sentences within the paragraph, a pitch reset mostly explained by the introduction of a new topic ([1], [2], [3], [4], [5], [6], [7], [8]). Authors also report a declination or downtrend throughout the paragraph, the last sentences being realized in a low and compressed register ([3], [9], [10], [11], [12], [13]). It appears in fact that the higher an element is connected in the text or in the discourse structure, the higher the register ([14],[15]). Register expansion is therefore associated with elements carrying new (or relevant) information, indicating topic change and usually positioned at the beginning of the structure. Register lowering is used for topic continuity, where elements within the lowering convey the same information. Register compression is found at the end of a topic discussed, hence found in final parts of the discourse. Without expectation then, sub-topics and parenthetical comments are reported to be associated with a compressed register ([16]).

Consequently, according to what has been found in the literature, we expect the clustering algorithm to detect a register reset at the beginning of a discourse structure or at the beginning of a new discussed issue, separating these units from the preceding ones in the clustering structure. We also suppose that the algorithm may detect a declination trend throughout the discussed issue, specifically at the end of the discourse structure, grouping together the units that are semantically linked. This last point which is currently under investigation, it will not be presented in this article. Such a function-oriented approach confronted with acoustic data may indeed allow the automatic extraction and prediction of functional information, hence contributing to the automatic mapping of prosodic form and function for speech synthesis.

2. Corpora

Four corpora were used in this study:

PAC (Phonologie de l'Anglais Contemporain, [36]) – A total of 30 minutes of newspaper article-like readings were selected from the PAC (5 female and 3 male speakers from Northern England; Lancashire, Greater Manchester and West Yorkshire).

AIX-MARSEC ([17]) **–** A total of 54 minutes of recording (15 female and 39 male speakers of standard British English) were selected from the AIX-MARSEC corpus. Mainly prepared monologues, the recordings correspond to commentaries, new broadcasts, lectures, religious broadcasts, magazine-style reporting, fiction, poetry, dialogues and propaganda.

PFC (Phonologie du Français Contemporain, [18]) **–** A total of 80 minutes of recording (6 female and 4 male speakers of regional French - Marseille) were selected from the PFC corpus. The recordings, being of three different speaking styles consist of newspaper article-like readings as well as guided and spontaneous conversations.

CID (Corpus of Interactional Data, [19]) **–** A total of 30 minutes of dialogue recording in a sound-proof room (3 female and 3 male speakers of regional French - Marseille) were selected from the CID corpus.

3. Measuring speakers' register

Detecting variations in speakers' register implies at first the efficient detection of their global shape, i.e. their global key and span. Many studies have used median or mean f0 to express register key and the difference between the extreme values or the standard deviation to express its span ([20], [21], [22], [23], [24], [25], [26]). However, since fo detection is very sensitive to microprosodic effects and octave errors, hence resulting in very error-prone results, many authors have

chosen, instead of the extrema, to limit their measure to quantiles, the difference between the 90th and 10th quantiles or the 95th and the 5th quantiles for example ([27], [28]). Therefore, an experiment was conducted ([29]) to test which quantiles give the best estimate of a speaker's register. First, manual annotation of maximum and minimum pitch of the AIX-MARSEC and PFC speakers were made from the editor window of the objects Sound and Pitch (the scale being adapted for each speaker's voice), thus dismissing microprosodic effects and octave errors in the analysis; then, a comparative study of different quantiles (from q05 to q95) was carried out to estimate which were best correlated with manual estimate of pitch extrema. In [30], the formulae 1.75*q75 and 0.75*q25 were taken to give best estimations of fo extrema while implemented as ceiling and floor (with +/- 10 Hz) in the MOMEL-INTSINT algorithm ([31], [32]). New investigation showed that q65*1.90 and q35*0.72 slightly improve the estimation of these extrema.

Comparing the detection of the manually annotated extrema (MMIN, MMAX; serving as reference) to the extrema values as obtained with the defined formulae (MIN35; MAX65) and to the extrema values as obtained in Praat (MINPRAAT, MAXPRAAT; i.e. by using the default floor of 75 and ceiling of 600 and then the functions Get minimum... and Get maximum...), it appears that the algorithm greatly improves the detection of speakers' register key and span (Figures1&2).

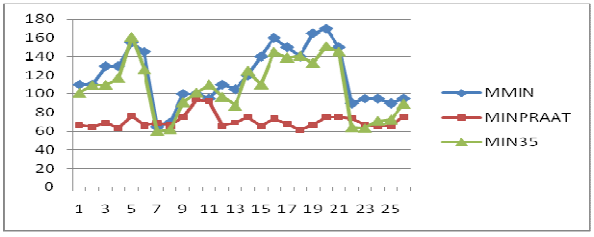

Figure 1: *Graphic representation of MMIN (given in Hertz) as compared to MINPRAAT and MIN35 as obtained from the PFC corpus (27 files selected).*

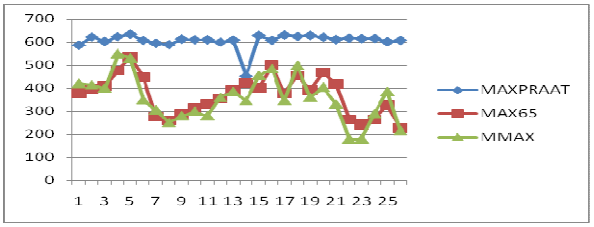

Figure 2: *Graphic representation of MMAX (given in Hertz) as compared to MAXPRAAT and MAX65 as obtained from the PFC corpus (27 files selected).*

4. Automatic detection of variations in register key and span

In [33], an improved version of our clustering algorithm, allowing the automatic detection of register variations, was proposed. First, the algorithm (1) recursively reduces the Euclidian distance between two consecutive tonal units in a space defined by key and span parameters (normalized values), and (2) calculates the difference between two consecutive units for key and span separately. Key is calculated in hertz then normalized with a logarithmic transformation. Span is calculated in semitones, thus normalized too. The formulae used in the calculation of the differences in key and span (DIFFKEY, DIFFSPAN) are given in (f1) and (f2).

(f1) sqrt(log2(median_unit)log2(median_prevUnit))^2

(f2) sqrt(log2(max/min_unit)–log2(max/min_prevUnit))2

Then, after obtaining consecutive differences between units, the clustering algorithm groups the units together according to their difference in key and span. The smaller the difference between two units, the sooner these units are branched together.

The output generated by the algorithm is a binary tree structure in the form of a layered icicle diagram (Figure 3). This representation allows the definition of the hierarchical structure and relational organisation of tonal units as reflected by register changes. Groups of units are therefore distinguished and an analysis of the distance between the leaf nodes according to key and span parameters (NDKEY & NDSPAN) allows boundary strength measurements between them. The larger the distance, the stronger the boundary between two groups. On the contrary, a short distance suggests that two consecutive units belong to the same group of units. The benefit of such an algorithm is that different units may be tested and that smaller units may be grouped together so as to indicate which unit may be under investigation, allowing "theory-neutral" analyses.

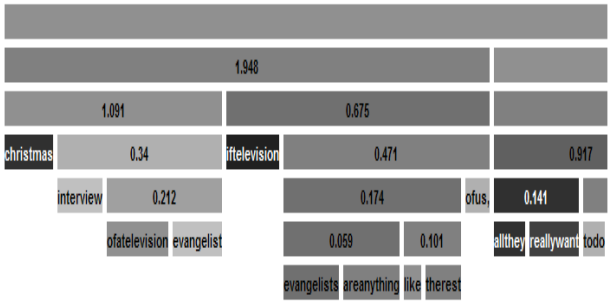

Figure 3: *Extract of a layered icicle diagram representation as obtained with the algorithm. The representation suggests that units "Christmas" to "evangelist" belong to the same group and are separated from the group of units starting with "if television". In fact, the distance between leaf nodes "evangelist" and "if television" indicates the presence of a strong boundary. The colour scale used indicates register key for each unit. The darker the colour, the higher the key.*

5. Detection and prediction of topic changes

5.1. Manual annotation of topic changes

To test the relevance of the algorithm, the topic structure of discourse was annotated in terms of topic change. The annotations were made by two labellers (the first author and a colleague) on 30 minutes of read speech from 8 English native speakers (PAC), 30 minutes of read speech from 10 French native speakers (PFC), 30 minutes of dialogue from 9 English-native speakers (AIX-MARSEC), and 30 minutes of dialogue from 6 French-native speakers (CID). Specifically, three levels of disjuncture between adjacent units were defined, a simplified version of Grosz & Sidner [34] as used in Fon [14] and Kong [35]. When a DSP0 is indicated between two units,

it means that these units share the same topic and the same relation to the units that dominate them, hence DSP0 indicates no discourse unit boundary. DSP1 stands for a hierarchically superior relation between two units, meaning that the units share related purposes, indicating, for example, a cause-consequence relation or again a clarifying relation. Finally, when two discourse units with no related discourse purposes or topics side together, a DSP2 boundary is inserted to indicate the introduction of a new issue. However, we already assume that further work is needed to refine this manual annotation.

5.2. Discourse structure and variations in register and pause duration

ANOVA analyses, carried out for the four corpora, showed a significant correlation between discourse structure and register changes. The DIFFKEY and NDKEY between two consecutive units, for DSP1 and DSP2 levels, significantly increases while approaching higher levels of disjuncture (pval < 2.2e-16) (Figure4), except for the CID data where only DSP2 is reported significant. DIFFSPAN is significatively correlated with DSP1 and DSP2 levels for the AM corpus (F(2,3447)=23.98, pval=4.549e-11); it is reported less significant than DIFFKEY for the PFC and the CID corpora (pval= 0.001995 & 0.02847 respectively) and not significant for the PAC Corpus (F(2,3003)=0.14, pval=0.8634). NDSPAN is highly correlated with DSP2 level (pval< 2e-16) while it is not with DSP1 level for the four corpora (pval>0.441). Thus, key appears as a stable parameter to indicate topic changes while span may be optional. Variations in span may be rather seen as marking a speaker's involvement while telling his/her story. It may be concluded that key and span parameters may convey different functions and have to be studied separately. Finally, ANOVAs analyses showed that pause duration is correlated with topic changes (pval < 2.2e-16), longer pauses being inserted before the introduction of a new issue.

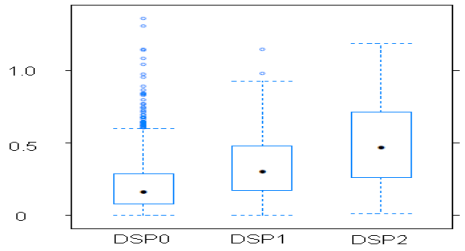

Figure 4: *Boxplots of difference in key depending on the three levels of discourse boundary, i.e. DSP0, DSP1 and DSP2 as obtained from the PFC data.*

5.3. Predicting topic changes

Having shown that discourse structure is significantly correlated with register key changes and pause duration, we investigate in this section the possibility of using these two parameters to predict the temporal location of discourse structure markers. We consider two classes, the first corresponding to units labelled DSP0 (class 1) and the second regrouping the DSP1 and DSP2 markers (class 2). The training of the binary classifiers is performed on the sample, for which the distribution proportion between the two classes is highly asymmetric (e.g. for the PFC, 93% of the units belong to class 1 and 7% to class 2). The first classifier takes as input parameter feature the pause duration, immediately preceding the unit under consideration. Training the classifier allows the definition of a pause duration threshold (from 0.49s to 0.79s, depending on the corpus) delimiting the two classes. The unit will be classified in class 1 if its feature value is lower than the threshold and in class 2 otherwise. The evaluation of the classifier is obtained on the sample by computing the confusion matrix (i.e. the counts of observed versus predicted classes on the sample). The scores of F-Measure associated with the prediction of discourse markers (class 2) are given for the four corpora in column 2, Table1.

Corpus	Pause	NDKEY	Both Features
PAC	0.321	0.181	0.325
PFC	0.766	0.298	0.754
AM	0.564	0.305	0.566
CID	0.472	0.232	0.468

Table 1: *Scores of F-Measure for the classifiers based on the pause duration feature (column 2), on the node distance feature (column 3) and on the combination of both features (column 4).*

The second classifier is based on the register key feature. The relevant quantity considered is herein the weighted node distance (i.e. the node distance weighted by difference in register key, NDKEY) between units obtained from the clustering tree structure. This classifier is less efficient than the one based on duration pause information as shown in Table 1, column 3.

Because of the low correlation previously obtained between the discourse structure and variations in span, register span feature has been excluded from classifying analyses.

Combining both pause and NDKEY features (Table1, column 4) reveals that adding register key information does not improve the prediction power of the classifier. It appears that pause information only is sufficient to predict topic changes, specifically for read speech in French.

6. Discussion

Classifiers have shown that pause duration is the best predictor for topic changes. However, it has to be noticed that the magnitude of the prediction power is corpus-dependant (from 0.32 to 0.76). Speakers may use different strategies to indicate a topic change. In reading task for example, while pause variations are widely used by French speakers, the case is not as straightforward for English ones.

Moreover and against all expectations, the addition of key feature has been reported as not sufficient enough to improve the prediction power. This may be explained by both the asymmetry of the class distribution and the supremacy of pause feature during the merging process. Again, speakers may vary their register key differently or use other prosodic cues to signal a topic change such as variations in speaking rate and intensity. It may be therefore interesting to merge these other prosodic features in the existing classifiers to improve their prediction power.

Furthermore, the objective detection of register variations as obtained with the clustering algorithm is correlated with manual annotation of topic changes. It may be assumed that the results obtained are dependant on the subjective annotation. The choice of three levels of structure boundary (DSP0, DSP1 & DSP2) may be questioned and considered not sufficient enough to understand and capture topic changes.

Finally, as the detection is objective, the algorithm may detect variations in register key and span that convey other functions

than topic changes (such as focus for example), functions which are not indicated in our annotation and therefore annotated with a DSP0 boundary if not linked to topic changes. This may have lowered the prediction power of the classifier.

7. Conclusions

If pause duration has been shown to be sufficient enough to predict topic changes, key feature has been reported as not bringing additional information as was expected by statistical analyses. However, the method consisting in using a clustering algorithm may be regarded as very promising, notably to better understand the hierarchical and organisational structure of the discourse, ie. how units are embedded in it and therefore improve the (automatic) representation of intonation patterns. For example, as already mentioned, the algorithm may be used for the analysis of declination phenomenon. Indeed, it may reveal the way declination is captured in phonological units such as the Accentual phrase or the Intonational Phrase.

8. References

[1] Lehiste, I., "Suprasegmentals", MIT., 1970

[2] Brazil, D., Coulthard, M. and Johns, C., "Discourse Intonation and Language Teaching", Longman, 1-82., 1980.

[3] Menn, L. & Boyce, S., "Fundamental frequency and discourse structure"., Language and Speech 25. 341-383., 1982.

[4] Pierrehumbert, J. and Hirschberg, J., "The Meaning of Intonational Contours in the Interpretation of Discourse"., In COHEN, Philip R.; MORGAN, Jerry; POLLACK, Martha E. (eds), Intentions in Communication, 1990, pp. 271-311Smith, J. O. and Abel, J. S., "Bark and ERB Bilinear Transforms", IEEE Trans. Speech and Audio Proc., 7(6):697-708, 1999.

[5] Arons, B., "Pitch-based Emphasis Detection for Segmenting Speech Recordings.", In PICSLP., September 18-22, Yokohama, Japan, (4): 1931-4., 1994.

[6] Kong, E., "The role of pitch range variation in the discourse structure and intonation structure of Korean", In *INTERSPEECH-2004*, 3017-3020.

[7] Mayer, J., Jasinskaja, E. and Kölsch, U., "Pitch range and pause duration as markers of discourse hierarchy: perception experiments", In *INTERSPEECH-2006*, paper 1290-Mon2FoP.7.

[8] Chiu-yu T., Chun-Hsiang C. and Zhao-yu S., "Investigating F0 reset and Range in relation to Fluent Speech Prosody Hierarchy"., Technical Acoustics 2005, (24) 279-285., 2005.

[9] Thorsen, N. "Text and intonation. A case study." In GREGERSEN, Kirstenin; BASBOLL, Hans (éds.), Nordic Prosody IV, Symposium on Prosody, Los Angeles, CA, USA, 4, 1986. Coll. Nordic Prosody, 4. Odense, Danemark : Odense University Press, 1987, p. 71-80 (coll. Odense Studies in Ling., 7).

[10] Hirschberg, J. and Pierrehumbert, J., "Intonational structuring of discourse"., In Proceedings of the 24th meeting of the Association for Computational Linguistics, 136-144, New York., 1986.

[11] Nakajima, S. and Allen, J.F., "Prosody as a cue for discourse structure", In *ICSLP-1992*, 425-428.

[12] Sluijter, A.M.C.; Terken, J.M.B., "Beyond Sentence Prosody: Paragraph Intonation In Dutch"., Phonetica, nov. 1993, pp. 180-188.

[13] Nicolas, P., Hirst, D.J., " Symbolic coding of higher level characteristics of fundamental frequency curves". In Proceedings of the 4th European Conference on Speech Communication and Technology, Madrid, 1995.

[14] Fon, J. "A cross-linguistic study and discourse boundary cues in spontaneous speech", Doctoral Dissertation, The Ohio State University., 2002.

[15] den Ouden, H., Noordman, L., and Terken, J. "Prosodic realizations of global and local structure and rhetorical relations in read aloud news reports" In Speech Communication, 51(2) February 2009, 116-129.

[16] Kutik, EJ., Cooper, WE. and Boyce S. "Declination of fundamental frequency in speakers' production of parenthetical and main clauses" In J. Acoust. Soc. Am. 73(5), 1731-1738., May 1983.

[17] Auran, C., Bouzon, C. & Hirst, D. " The Aix-MARSEC Project: An Evolutive Database of Spoken British English", in Speech Prosody 2004, ISCA.

[18] Delais-Roussarie, E. & Durand, J. "Corpus et variation en phonologie du français: méthodes et analyses", Presses Universitaires du Mirail., 2003.

[19] Bertrand, R., Blache, P., Espesser, R., Ferré, G., Meunier, C., Priego-Valverde, B. and Rauzy, S. "Le CID-Corpus of Interactional Data-: protocoles, conventions, annotations" In Travaux Interdisciplinaires du Laboratoire Parole et Langage d'Aix en Provence, 25, 25-55., 2007.

[20] Huttar, G. "Relations between prosodic variables and emotions in normal American English utterances", In Journal of Speech and Hearing Research 11. 481-7., 1968.

[21] Jassem, W. "Pitch and compass of the speaking voice." In Journal of the International Phonetic Association 1. 59-68., 1971.

[22] Van Bezooijen, R. "Sociocultural aspects of pitch differences between Japanese and Dutch women" In Language and Speech, 38(3): 253-265., 1995.

[23] Clark, R. "Using prosodic structure to improve pitch range variation in text to speech synthesis.", In Proc. XIVth international congress of phonetic sciences, volume 1, pages 69-72, 1999.

[24] Mozziconacci, S.J.L. "Speech variability and emotion: production and perception," Ph.D. thesis, Eindhoven, The Netherlands., 1998.

[25] Lennes, M. & Anttila, H. " Prosodic features associated with the distribution of turns in Finnish informal dialogues"., In: Korhonen, P. (ed.) The Phonetics Symposium 2002. Helsinki University of Technology, Laboratory of Acoustics and Audio Signal Processing, Report 67, 149-158.

[26] Rietveld, T., and Vermillion, P., "Cues for perceived pitch register", In *EUROSPEECH-2001*, 1399-1402.

[27] Patterson, D & Ladd, R. "Pitch Range Modelling: Linguistic dimensions of variation" in ICPHS99. 1169-72., 1999.

[28] Patterson, D. " A linguistic approach to pitch range modeling". PhD dissertation, University of Edinburgh, 2000.

[29] De Looze, C. In progress. Analyse et interprétation de l'empan temporel des variations prosodiques en français et en anglais contemporain. Doctoral thesis, Université de Provence.

[30] De Looze,C. and Hirst, DJ., "Detecting changes in key and range for the automatic modelling and coding of intonation"., In Speech Prosody 2008, Campinas, Brazil.

[31] Hirst, D.J., "A Praat Plugin for MOMEL and INTSINT with improved algorithms for modelling and coding of intonation.", In *Proc*. Int. Conf. Phonetic Sci. XVI, Saarbrucken., 2007.

[32] Boersma, P. and Weenink, D. Praat: doing phonetics by computer (Version 4.6.35) [Computer program]. Downloadable from http://www.praat.org/. 2007.

[33] De Looze, C., "Automatic detection of register changes for the analysis of discourse structure", PaPI2009.

[34] Grosz, B. G. and Sidner , C.L. "Attention, intentions and the structure of discourse", Computational Linguistics, 12:175-204., 1986.

[35] Kong, E. " The role of Pitch Range Variation in the Discourse Structure and Intonation Structure of Korean", In InterSpeech, 8th ICSLP., 2004.

[36] Carr, P. Durand, J. La Phonologie de l'Anglais Contemporain : usages, variétés et structure : The Phonology of Contemporary English: usages, varieties and structure., 2003.