Prosodic phrasing evaluation: measures and tools

Klim Peshkov, Laurent Prévot, Roxane Bertrand

Aix-Marseille Université
Laboratoire Parole et Langage
5 avenue Pasteur
Aix-en-Provence, France
klim.peshkov@lpl-aix.fr, laurent.prevot@lpl-aix.fr

Abstract

Over the recent years several transcripion systems and tools have been created for marking prosodic phrasing. Although they correspond to different theoretical stances and objectives, it seems important to us to be able to compare the results of the tools and to study the reliability of the coding systems. However, only a few studies [0], [1] have focussed on reliability. We compare several segmentation evaluation metrics as well as intercoder reliability measures. About evaluation metrics, methodologies are coming mostly from clause or word segmentation: (i) precision and recall on boundaries ; (ii) WindowDiff and (iii) segmentation similarity. With regard to intercoder agreement, we discuss the standard measure (κ) and how it is applied to segmentation tasks. The poster consists in a practical application to two cases: (i) an evaluation of prosodic tools and (ii) a reliability evaluation of annotation campaign.

Index Terms: evaluation; intercoder reliability; speech prosody; prosodic phrasing detection

1. Introduction

Prosodic information is useful to answer linguistic questions and to create applications which deal with speech. Annotating prosody of large corpora by human means is costly and rarely possible. Automatic tools have been created to automate detection of prosodic events, but in order to use them, we would like to have a better idea of their performance. In order to evaluate tools in terms of human performance, one has to rely on reference segmentation made by human. However, using annotation made by only one person is risky, because a part of answers might be simple guesses. With multiple annotators, it is possible to create highly reliable "gold standard" [2]. First step in this direction is to obtain interannotator agreement measure.

The evaluations presented below are performed on the Corpus of Interactional Data (CID) [3]. This is a corpus made of 8 conversations of one hour involving two speakers. The protocol for obtaining this data was made in such a way that the interactions are highly natural featuring a lot of overlap and disfluencies.

Section 2 discusses precision/recall metrics and WindowDiff metrics in application to evaluation of prosodic phrasing tools evaluation. Section 3 presents estimation of interannotator agreement for prosodic phrasing annotation using κ statistics.

2. Evaluation of automatic segmentations

A number of tools for automating prosodic analyses have been proposed for French. We can cite Analor [6], Momel-Intsint [7], Prosogram [8]. Among these tools only Analor is directly conserned with prosodic phrasing. We also implemented an algorithm proposed by Simon et Degand (henceforth DS) [9], which is based on phonetic cues such as syllable length and fundamental frequency variation. Our baseline segmentation in Inter-Pausal Units (IPU), which assigns boundaries before and after pauses longer than 200 milliseconds.

In order get a more precise idea about different tools for prosodic phrasing detection, we want to compare quantitatively the outputs of these tools with reference manual annotation and also to compare different outputs of the tools between them. In this section, we use an annotation of intonation phrases (IP) made by one expert linguist as reference segmentation.

2.1. Precision, recall and f-measure

Precision, recall and f-measure are conventional evaluation metrics from information retrieval. Applied to segmentation task, separate measures for left boundaries, right boundaries and the entire units. This method was used, for example, for the shared task of CoNLL-2001 (Conference on Computational Natural Language Learning) [4]. In our case, we do not work with text, but with aligned transcripts. Hence the alignment is not always perfect. We adopted a delta of 160 ms to tolerate near small mismatches. The value corresponds to the average lenghth of syllables in our corpus.

Table 1 presents results of evaluation of tool's outputs using expert annotation. Low rates of detection, especially in case of the whole units, may be due to the fact, that the tools and the manual segmentation contain prosodic objects of different levels. The DS algorithm shows the best results in the detection of starts, ends and whole units. All the tools are better at detection of starts of the units than their ends.

2.2. WindowDiff

It should be noted that, when used for segmentation evaluation, information retrieval metrics present a serious drawback. They do not take in consideration the distance between the borders of the segmentations being compared. Near-miss errors are penalized as heavily as insertion or deletion of borders and using delta can result in a bias.

WindowDiff metrics was introduced to adress this problem

		spk1			spk2			Mean		
		Prec.	Recall	f	Prec.	Recall	f	Prec.	Recall	f
S	IPU	82.6	39.2	53.2	83.3	43.0	56.7	83.0	41.1	55.0
	DS	77.9	44.5	56.6	78.2	49.5	60.6	78.0	47.0	58.6
	An.	82.1	34.3	48.4	84.9	35.6	50.2	83.5	35.0	49.3
E	IPU	72.1	34.3	46.5	74.9	38.6	51.0	73.5	36.4	48.7
	DS	67.7	38.7	49.2	69.4	44.0	53.8	68.6	41.3	51.5
	An	76.0	31.8	44.9	81.2	34.1	48.0	78.6	32.9	46.4
U	IPU	30.2	14.4	19.5	37.6	19.4	25.6	33.9	16.9	22.5
	DS	30.7	17.5	22.3	36.8	23.3	28.6	33.7	20.4	25.4
	An.	30.5	12.8	18.0	38.9	16.3	23.0	34.7	14.5	20.5

Table 1: Precision and recall. Evalutation of segmentations in terms of human preformance

[5]. The algorithm operates as follows. It constists in moving a fixed-length window along the two segmentations (cf. Figure 1[1]), one unit at a time. On the scheme, the length of the window, which is represented by arrows, is 5 units. For each position, the algorithm compares the numbers of borders in both segmentations. If the number of borders is not equal, the difference of the numbers is added to the evaluated algorithm's penalty. The sum of penalties is then divided by the number of stops, yielding a score between 0 and 1. The score 0 means that the segmentations are identical. The length of the window is set to $1/2$ of the average length of a unit in the reference segmentation.

Initially, WindowDiff was created for text segmentation tasks. When applying it to prosodic units evaluation in time-aligned transcripts, we had to adapt it to our case by indroducing a time-based step. If we had chosen to move the window by unit-based step, we would loose time dimention of our data. That's why we introduced a time-based step to move the window. Setting shorter step provides higher resolution of evaluation (but requires more computation time). Results shown here were obtained with a step of 20 milliseconds.

One of the problems of this relatively new metrics is that it is difficult to interpred results in absolute terms. In order to have the first picture of WindowDiff's behaviour, we tested it by perturbating identical segmentations. Figure 2 presents the evolution of WindowDiff score (y-axis) depending on the proportion of randomly moved boundaries (average distance of perturbation is 2.6 seconds and the minimal distance is set to 100 milliseconds). The score evolves in a linear fashion, but quite slowly. When 99% of the boundaries are moved, it reaches only 0.5. The score 0.3 can be interpreted as high divergence between segmentations, because it corresponds to 50% of moved boundaries.

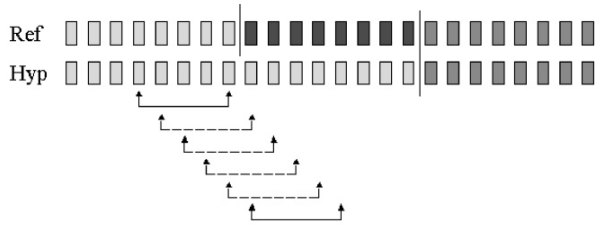

Figure 1: *WindowDiff metrics*

Table 2 presents WindowDiff metrics of tools' outputs in comparison to manual annotation. All the results indicate high divergence with the manual annotation. As in the case of precision and recall metrics, the DS algorithm's segmentation is

[1] reproduced from [5].

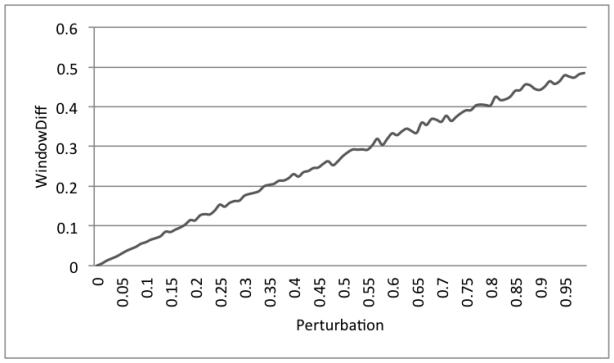

Figure 2: *WindowDiff test by boundaries perturbation*

	spk1	spk2
IPU	0.275	0.281
DS	0.263	0.265
An.	0.306	0.321

Table 2: WindowDiff. Evaluation against expert annotation

the closest to the reference. Although there is only a slight improvement over the baseline.

The next set of results (Table 3) is a comparison between automatic segmentations. The comparison was made in both directions, because depending on the choice of reference segmentation, the length of the window changes, producing different results. This is why, the results of IPU-Analor and Analor-IPU differ. It follows from the table, that DS algorithm is very close to IPU, and Analor's outputs differ a lot from both.

3. Interannotator agreement

During an annotation campaign of prosodic phrasing by naive annotators, the annotators were asked to assign a number between 0 and 4 to words' right boundaries, corresponding to 4 levels of prosodic break (similar to break indices in the ToBI system). 0 is the default boundary between two words without prosodic marking. Thus, each word's right boundary represents a decision point. All 8 dialogues of the CID corpus were annotated, each speaker was annotated by two judges.

In order to obtain a rough evaluation of the reliability of annotations we used a simple inter-annotator measure, the κ statistics. It is interpreted as "the proportion of joint judgments in which there is agreement, after chance agreement is excluded" [10]. The value of κ ranges between -1 and 1.

Table 4 shows interannotator reliability for several speakers of our corpus. First line takes in consideration all the four levels. The agreement is low, which means that the task was too difficult for the annotators. Second and third lines flatten levels to arrive to higher scores using just 2 classes instead of 4.

	Reference segmentation		
	IPU	An.	DS
IPU	–	0.311	0.077
An.	0.158	–	0.233
DS	0.089	0.390	–

Table 3: WindowDiff. Comparison between tools' segmentations

	spk1	spk2	spk3	spk4	spk5	spk6	Mean
0 1 2 3	0.38	0.28	0.27	0.48	0.16	0.36	0.32
(0 1) (2 3)	0.45	0.41	0.31	0.62	0.17	0.46	0.40
0 (1 2 3)	0.58	0.52	0.56	0.70	0.27	0.66	0.55

Table 4: Interannotator agreement

4. Conclusions and future work

Above, we presented such evaluation metrics as (i) precision and recall and (ii) WindowDiff with examples of their usage in the context of evaluation of tools for prosodic phrasing detection. The interannotator agreement of prosodic boundaries was also discussed with an example of results.

We continue to experiment with Analor tool by tweaking its parameters with the aim to obtain segmentations which would be more similar to the IPs.

In future, we would like to experiment with segmentation similarity metrics. It was proposed by [11] as an improvement of WindowDiff. This metrics relies on edit distance between the boundaries to compute penalties.

Acknowledgements

The author would like to thank Provence-Alpes-Cte d'Azur region which supported this work.

5. References

[0] Lacheret, A. and Obin, N. and Avanzi, M. "Design and evaluation of shared prosodic annotation for spontaneous French speech: from expert knowledge to non-expert annotation" Proceedings of the Fourth Linguistic Annotation Workshop: 265-274, 2010

[1] Breen, M. and Dilley, L.C. and Kraemer, J. and Gibson, E. "Inter-transcriber reliability for two systems of prosodic annotation: ToBI (Tones and Break Indices) and RaP (Rhythm and Pitch)" In press, 2013

[2] Beigman Klebanov, B. and Beigman, E. "From Annotator Agreement to Noise Models" Computational Linguistics, 35(4):495-503, 2009

[3] Bertrand R. and Blache, P. and Espesser, R. and Ferr, G. and Meunier, C. and Priego-Valverde, B. and Rauzy, S. "Le CID — Corpus of Interactional Data — Annotation et Exploitation Multimodale de Parole Conversationnelle" Traitement Automatique des Langues 49(3):1-30, 2008

[4] Tjong, E.F. and Sang, K. and Djean, H. "Introduction to the CoNLL-2001 shared task: clause identification", Proceedings of the 2001 workshop on Computational Natural Language Learning, 7:127-132, 2001

[5] Pevzner, L. and Hearst, M. A. "A critique and improvement of an evaluation metric for text segmentation", Computational Linguistics, 28(1):19–36, 2002

[6] Avanzi, M. and Lacheret-Dujour, A. and Victorri, B. "A corpus-based learning method for prominence detection in spontaneous speech" Vth International Conference Speech Prosody, 2010

[7] Hirst, D. "A Praat plugin for Momel and INTSINT with improved algorithms for modelling and coding intonation" Proceedings of the XVIth International Conference of Phonetic Sciences: 12331236, 2007

[8] Mertens, P. "The prosogram: Semi-automatic transcription of prosody based on a tonal perception model" Proceedings of Speech prosody, 2004

[9] Simon, A. C. and Degand, L. "On identifying basic discourse units in speech: theoretical and empirical issues" Discours, 4, 2009

[10] Cohen, J. "A coefficient of agreement for nominal scales" Educational and psychological measurement, 20(1):37-46, 1960

[11] Fournier, C. and Inkpen, D. "Segmentation similarity and agreement" Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: 152-161