On the use of probabilistic grammars in speech annotation and segmentation tasks

Irina Nesterenko, Stéphane Rauzy

On the use of probabilistic grammars in speech annotation and segmentation tasks

Irina Nesterenko, Stéphane Rauzy
Université de Provence, Laboratoire Parole et Langage

Abstract

The present paper explores the issue of corpus prosodic parsing in terms of prosodic words. This question is of importance in both speech processing and corpus annotation studies. We propose a method grounded on both statistical and symbolic (phonological) representations of tonal phenomena and we have recourse to probabilistic grammars, whithin which we implement a minimal prosodic hierarchical structure. Both the stages of probabilistic grammar building and its testing in prediction are explored and quantively and qualitatively evaluated.

1. Introduction

Current linguistic research is largely founded on annotated speech corpora: among other things, such a demarche is adopted to collect evidence of how language systems function in the speech production and speech perception. Linguistic annotation of a speech corpus is in itself the process of enriching the row data with additional analytic notations. The latter cover various domains of language analysis (phonetics, phonology, prosody, semantics, syntax, pragmatics, etc.) in the line with a multi-disciplinary approach to language phenomena. Consequently, much work is looking for automatic and semi-automatic methods to annotate the row data on different levels.

Numerous formal representations have recently been proposed for different linguistic levels. Moreover, several components of grammar operate with their own hierarchical representations: this is namely the case of syntax and prosody. The issue of prosodic structure and prosodic phrasing is central to our present research. Prosodic phrasing has received a formal treatment within the framework of prosodic phonology [1]. Within prosodic phonology framework, it is assumed that each level of prosodic hierarchy stands for an interface between phonological module and another module of the grammar. At the same time prosodic structure controls for the acoustic organisation of the spoken material and consequently functions as a pivot structure in dealing with multiple interfaces.

To annotate prosodic phrasing phenomena means to establish the distribution of prosodic boundaries of different strengths. Consequently, there is a quest for heuristics, which allow for discriminating between different levels of prosodic constituency. In this respect, the issue of (semi-automatic) annotation of prosodic phrasing phenomena is intimately related to the problematic of speech segmentation and speech possessing. In fact, prosodic breaks divide the speech flow into smaller units, thus facilitating the processing. The psycholinguists are particularly interested in how the listeners segment the speech flow into the word-sized chunks in order to recognize the individual words and ultimately understand the meaning of the utterance. It was established that this process is largely founded on fine-grained phonetic information related to different levels of linguistic organisation, e.g. allophonic variation, phonotactic patterns [2], matrical constraints [3-4], durational and tonal patterns [5-6]. The cues to the speech segmentation are thus redundant. At the same time, the scientific community agrees that such cues do not determine the precise locations of word boundaries, but rather indicate the possible boundary locations.

In our study grounded on a corpus of Russian spontaneous speech, we explore the issue of corpus parsing in terms of prosodic words and we propose a method grounded on both statistical and symbolic (phonological) representations of tonal phenomena. Note that the role of tonal phenomena in word segmentation has not been largely studied for Russian. The rest of this paper is organised as follows: section 2 discusses the prosodic model underlying our empirical study and subsequent mathematical modelling. The goal of section 3 is to specify a corpus our study is based on and its preliminary annotation as well as the mathematical apparatus we have recourse to in our study.

Section 5 presents the results for both grammar building stage and its evaluation in prediction. Finally we discuss the impact of the approach adopted and required future work.

2. Prosodic representation

In modern conceptions, prosody is interpreted as an organisational system [7] that could be exhaustively specified via the analysis of tonal and rhythmical layers as well as that of prosodic phrasing.

Different annotation schemes have been proposed for the tonal component of prosody [8-9]. Our study is based on MOMEL-INTSINT prosodic annotation protocol [10, 11] which was recently coupled with the IF (for *Intonation Functions*) functional annotation system [12]. We assume a prosodic representation which distinguishes three intermediate levels between the acoustic signal and the level of prosodic functions, these levels being an underlying phonological representation, a surface phonological representation and a phonetic representation. The overall architecture developed for the tonal component of prosody is subject to the interpretability constraint, which stipulates that representations proposed at each level should be interpretable at both adjacent levels. The interpretability constraint follows from the considerations of the role of phonological representations: we assume that a phonological representation must provide the information necessary both for the pronunciation of the utterance and for its syntactic and semantic interpretation.

For the purposes of present research, we further discuss the details of phonetic and surface phonological representations.

At the level of phonetic representation, two components are factored out from the f0 curve [13], a macroprosodic component and a microprosodic component. The first corresponds to a continuous smooth intonation curve, tightly associated with the prosodic meaning, while the second answers for the deviations from the smooth curve caused by the nature of the current segment. The macroprosodic component is further modelled via the application of the MOMEL algorithm [14]. This modelling is grounded on the definition of target points in time ~ frequency space: these target points correspond to the inflections in f0 curve where the slope is null (i.e. the first derivative equals zero). To obtain a smooth curve, the pitch targets are linked by a quadratic spline function. The level of phonetic representation is primarily acoustically oriented. At the same time, the tonal targets could be viewed as the sites where the speaker voluntary changes the direction of the fundamental frequency curve to achieve his/her communicational goals. The MOMEL algorithm is currently implemented under the Praat software [15].

At the upper, surface phonological level, the f0 targets receive a symbolic coding in terms of the Intsint prosodic alphabet [10]. The INTSINT alphabet comprises 8 distinct symbols. In this annotation the target points are characterised either globally with respect to the speaker's pitch range (via the long-term parameters of key and range; the corresponding labels are T(op), B(ottom) and M(edium)) or locally, by the reference to the preceding target (H(igher), L(ower), S(ame)). The H and L labels have the iterative variants D(ownstepped) and U(pstepped). The Insint coding of detected target points provides a symbolic representation of tonal phenomena, which underlines probabilistic modelling proposed in our study.

3. Empirical study design

3.1. Corpus and annotations

Our study is based on a corpus of Russian spontaneous speech. This corpus was collected for the INTAS project 915 at the department of Phonetics, Saint-Petersburg State University. For the current study, the recordings of an informal spontaneous dialogue between two female speakers in their twenties were used and the productions of one of the speakers were analysed (17 minutes of speech including pauses).

As we stipulated earlier, the phonological representation of the pitch phenomena used in our study is the one obtained via the application of MOMEL-INTSINT algorithm. At the same time, we integrated into our model a minimal hierarchical structure. MOMEL-INTSINT algorithm crucially relies on two speaker dependent parameters: *key* and *range* (span), which define together the speaker's pitch range. Pitch range variations being rightly related to the message informational organisation, we choose to apply this annotation to the units corresponding to one speaker's turn in the dialogue. At the same time, our study explores the issue of corpus and speech parsing in terms of prosodic words, which correspond to the lower level in our prosodic annotation.

Next, the relationship of prosodic words to lexical words needs be clarified. In fact, Nespor and Vogel [1] distinguish two levels of prosodic phrasing, the level of prosodic words and that of clitic groups: if the former are homologous to lexical words, the latter are formed by a content word and all its clitics. However, other constituencies either include only one unit or propose a recursive structure. Simultaneously, studying the role of tonal cues to segmentation, we should take into account the distribution of prosodically meaningful tonal events, namely, pitch accent distribution. A set of units, associated with a pitch prominence have been proposed within intonational phonology: accentual units, tone association domains, tone units of the British school, feet and prosodic words. But for most of these units, their relevance for the prosodic analysis of Russian was not proved by experimental studies. Hence, we decided to focus our study on prosodic words: Russian tradition of prosodic studies relies on prosodic words in the description of rhythmical structure of speech. Note as well that speech production studies [16] argue that at the stage of preliminary processing, the unit of phonological encoding is the prosodic word.

The corpus was manually annotated as to the distribution of prosodic word boundaries and further each prosodic word received its annotation according to whether the word bears or not a pitch prominence. The corpus comprises 825 prosodic words.

The corpus annotation procedure allowed us to obtain a minimal hierarchical structure, which comprises two levels: the level of speaker's turns in conversation and the level of prosodic words. Next, we extracted tonal patterns, which are associated with each prosodic word. The annotation has been done in Praat and further transformed into XML format (cf. fig. 1). This hierarchical symbolic representation is further used as an input for the probabilistic grammar building module to infer the regularities about Intsint labels phonotactics. Next section presents a mathematical apparatus used in our study.

```
<block>
        <P>
                <U />
                <H />
                <B />
        </P>
        <N>
                <S />
                <S />
        </N>
         <N>
                <H />
        </N>
</block>
```

Fig. 1: Example of XML corpus annotation

3.2. Mathematical apparatus

One of the central assumptions of the present study postulates that there are the regularities in the Intsint tonal patterns and consequently the dependency relations between the labels. Our modelling relies on the apparatus of probabilistic grammars. Two mathematical concepts are of interest: the concept of conditional probability and that of entropy.

Consider the general case when we dispose of N categories c_i to annotate a speech phenomenon. We assume as well to dispose of a training speech corpora, from which the distributions of tonal categories as well as their interdependencies are studied. Our task is then to estimate the probability of any produced sequence, say for example the time-ordered sequence (c_1, c_2, c_3, c_4) which corresponds to the f0 curve annotated in terms of INTSINT tonal categories as (U, S, T, D). The probability of this sequence is calculated with the application of the concept of conditional probabilities:

$$P(c_1, c_2, c_3, c_4) = \pi_1 * \pi_2 * \pi_3 * \pi_4,$$

where $\pi_1 = P(c_1)$, $\pi_2 = P(c_2| c_1)$, $\pi_3 = P(c_3| c_1, c_2)$ and $\pi_4 = P(c_4| c_1, c_2, c_3)$. Herein the quantity $\pi_3 = P(c_3| c_1, c_2)$ stands for the conditional probability of the category c_3 given the preceding sequence of tonal labels (c_1, c_2).

In our study a Patterns model [17] is tested. The Patterns model is a new method belonging to the family of probabilistic finite state automaton approaches like n-gram models, for example (see [18] for a presentation of Hidden Markov Models). The Patterns model is characterized by an optimal extraction of the information content contained in the training corpora. Contrary to the n-gram model, the left context is not limited to a fixed number of symbols but rather takes into account the regularities observed in the corpus: if a sequence of tonal labels frequently occurs in the training database, the model calculates and memorises conditional probabilities for all the categories given the pattern. Three Patterns models were built:

- Flat model, for which no hierarchical structure is specified;
- Hierarchical model with two levels distinguished;
- Hierarchy & Prominence model, which integrates both the hierarchical structure and the distinction between prominence bearing and non-prominent prosodic words.

To evaluate the performance of the model, we resort to the measure of entropy, which is the measure of the informational organisation of the system. For a tonal unit the entropy allows us to measure the informational charge of this unit and consequently to answer the question of how informative this unit is. Simultaneously, for a given distribution it quantifies the difference between this distribution and an equiprobable distribution of the categories. The entropy of the system varies between *0* and *lnN* (where *N* is the number of categories of the encoding scheme): an entropy of 0 characterises a completely deterministic system, while an entropy of lnN is found for the equiprobable distribution. We will introduce the concept of normalised entropy to bring the entropy values to the interval between 0 and 1. Subsequently, to evaluate the performance of the patterns model we calculate the entropy of the probability distribution with and without the language model.

When the probabilistic models of the label sequences were built we sought to test them in prediction by applying the Viterbi algorithm: given the sequence of tonal labels and the underlying probabilistic grammar, we look for the optimal distribution of the prosodic words' boundaries. The performance of the prediction heuristics is quantified and evaluated with the measures of recall, precision and F-measure, traditionally used in information retrieval studies [18].

4. Results

4.1. Model building stage

The table 1 presents the values of entropy and normalised entropy for three experimental conditions (Flat, Hierarchical and Hierarchy&Prominence models). WithoutM values stand for the measures of information organisation of the frequency distribution of tonal labels, when no Patterns model is specified: we evaluate so the information weight of the frequency distribution alone. The values of normalised data are close to 1, which points out at the quasi-equiprobable distribution. The smallest value of normalised entropy for the WithoutM situation is observed for the Hierarachical Model, which probably indicates that the frequency distribution of the tonal labels bears more information load.

Our data show that WithoutM and WithM entropy values differ for three experimental conditions: to take into account one of the probabilistic models of the frequency distributions for the Intsint tonal labels reduces considerably the normalised entropy of the model. Moreover, a conditioning effect is more marked if a rudimentary hierarchical model is taken into account.

These data are particularly interesting in the context of the debate as to the relevance and meaningfulness of two prosodic annotation systems, ToBI, a standard in intonational phonology

framework [8], and INTSINT. The authors of ToBI criticize the INTSINT notation for just fixing the alternations of rises and falls without establishing any regularity for anchoring tune to text. In fact, as we pointed out earlier, the INTSINT annotation should be coupled with IF functional annotation in order to analyse a part of pitch contour as a contrastive pitch event. At the same time, our results point out that there do are the regularities in the phonotactics of INTSINT pitch labels, which combine to produce prominence landing pitch movements. Moreover, in our study we integrated only a part of IF annotation (namely, the distinction between prominent and non-prominent units). We could expect that the values of normalised entropy will be smaller if a more precise functional annotation be incorporated in our probabilistic grammar. This work should nevertheless be grounded on a larger speech corpus.

Table 1. Entropy and normalised entropy for three probabilistic models built

Model	Entropy		Normalised Entropy	
	WithoutM	WithM	WithoutM	WithM
Flat model	2.259	1.796	0.942	0.749
Hierarchical model	2.064	1.494	0.897	0.649
Hierarchy&Prominence model	2.696	1.638	0.915	0.556

4.1. Testing models in prediction

On the next step, we evaluated the performance of two probabilistic models, Hierarchical model and Hierarchy&Prominence model, in prediction of the distribution of prosodic words boundaries. As we specified early, the prediction heuristics uses Viterbi algorithm and so, tests different solutions with a boundary inserted after every Intsint symbol. The algorithm looks for the optimal distribution of prosodic word boundaries within the speaker's turn in conversation, i.e. within the unit of higher level in the hierarchical structure. The confusion matrix and the corresponding evaluation statistics are presented in tables 2-3.

Table 2. Confusion tables for Hierarchical and Hierarchy & Prominence models

Hierarchical Model

	Prediction	
	no boundary	boundary
no boundary	2003	190
boundary	329	497

Hierarchy & Prominence model

	Prediction	
	no boundary	boundary
no boundary	2003	194
boundary	398	428

Table 3. Evaluation metrics

	Hierarchical Model	Hierarchy&Prominence Model
Precision	0.72	0.688
Recall	0.6	0.518
F-measure	0.655	0.591

The recall quantifies the accuracy of the model, i.e. the proportion of correctly predicted cases over the data: the overall accuracy of the algorithm is 0.6 in case of Hierarchical model and 0.52 for Hierarchy&Prominence model. On the other hand, the precision measures the proportion of correctly predicted locations of prosodic word boundaries over all the boundaries, which were predicted. The precision values are greater: it means that when a boundary is inserted, in 70% of cases it is inserted at the right location. For the optimal performance of the model, the couple <recall, precision > should show maximal values: this is the f-measure statistics which take into account simultaneously the values of precision and recall. In our study, Hierarchical model benefits from an f-measure of 65.5% and overrules the Hierarchy&Prominence model (59.1%). We can

conclude that there is plenty of room for the improvement, though it was an interesting goal to test the probabilistic grammar techniques in speech segmentation task and the overall quality of prediction is comparatively good.

5. Discussion and conclusions

Methodologically oriented, the present work sketches an approach for prosodic information retrieval and speech segmentation, based on both symbolic and probabilistic information.

The goal of our study was three-fold:

- to test whether there are the regularities in Intsint labels sequences associated with prosodic words;
- to model how the probabilities over the tonal space could be integrated into speech segmentation task, performed by both human listeners and automatic corpus segmentation heuristics;
- to test an impact of a rudimentary hierarchical structure.

Our data confirm that there are the probabilistic regularities in the way how the Intsint labels are combined. We hypothesise that this information is explored by listeners: in fact, psycholinguistic studies have underscore the role of frequency and probabilities information at the early stages of speech signal processing.

The proposed algorithm was tested in prediction of prosodic words boundaries. The model achieves 60% of correct prediction and benefits from the f-measure of 65.5%: the overall quality of prediction is so comparatively good. Note that our model was trained on a very limited speech corpus: we expect that to dispose of more annotated data would help refine the probabilistic model and augment the quality of prediction heuristics.

We constated as well that to implement a hierarchical model within a probabilistic grammar contributes to the information organisation within the tonal space. Note that the model applied in our study does not reflect a more developed prosodic constituency. A more fine-grained annotation of prosodic functions could be also implemented if more data are available. At the same time, the findings of our present study encourage us to pursue the application of probabilistic methods in the linguistic research, both for speech annotation task and in treatment of more theoretically oriented problematic of speech processing.

References

1. *Nespor, M., Vogel, I.* Prosodic Phonology, Foris Publication, Dordrecht, 1986.
2. *Weber, A.* The role of phonotactics in the segmentation of native and non-native continuous speech. Proceedings of the Workshop on Spoken Access Processes, 143-146, 2000.
3. *Cutler, A., Mehler, J., Norris, D. and Segui, J.* The syllable's differing role in the segmentation of French and English. Journal of Memory and Language, Vol. 25, 1986, p. 385-400.
4. *Cutler, A., Otake, T.* Mora or phoneme? Further evidence for language-specific listening. Journal of Experimental Psychology: Human perception and performance, Vol. 14, 1994, p. 113-121.
5. *Turk, A.E., Shattuck-Hufnagel, S.* Word-Boundary-Related Duration Patterns in English. Journal of Phonetics, Vol. 28, 2000, p. 397-440.
6. *Di Cristo, A.* Vers une modélisation de l'accentuation du français : seconde partie. French Language Studies, Vol. 10, 2000, 27-44.
7. *Beckman, M.E.* The parsing of prosody. Language and Cognitive Processes, Vol. 11, 1996, 17-67.
8. *Beckman, M.E., Hirschberg, J., Shattuck-Hufnagel, S.* The original ToBI system and the evolution of the ToBI framework. In J*un, Sun-Ah* (ed.), Prosodic Typology: The Phonology of Intonation and Phrasing. Oxford: Oxford University Press, 2005.
9. *Gussenhoven, C.* Transcription of Dutch Intonation. In *Jun, Sun-Ah* (ed.), Prosodic Typology: The Phonology of Intonation and Phrasing. Oxford: Oxford University Press, 118-145, 2005.

10. *Hirst, D., Di Cristo, A.* Intonation Systems. A Survey of Twenty Languages. Cambridge, Grande Bretagne : Cambridge University Press, 1998.
11. *Hirst, D., Di Cristo, A., Espesser, R.* Levels of description and levels of representation in the analysis of intonation. In M. Horne (ed.), *Prosody: Theory and Experiment*, Dordrecht : Kluwer Academic Press, 51-87, 2000.
12. *Hirst, D.* Form and function in the representation of speech prosody. Speech Communication (Special Issue), Vol. 46(3-4), 2005, p. 334-347.
13. *Di Cristo, A., Hirst, D.* Modelling French micromelody: analysis and synthesis. Phonetica, 43 (1):11-30, 1986
14. *Hirst, D., Espesser, R.* Automatic modelling of fundamental frequency using a quadratic spline function. TIPA 15, 1993, p. 71-85.
15. *Boersma, P., Weenink, D.* Praat : a system for doing phonetics by computer. 1995-2007. Available from <http://www.fon.hum.uva.nl/praat/>
16. *Levelt, W. J. M.* Accessing words in speech production: stages, processes and representations. Cognition, 42, 1992, 1-22.
17. *Blache P., Rauzy, S.* Mécanismes de contrôle pour l'analyse en Grammaires de Propriétés. Proceedings of the TALN Conference, 2006, p. 415-424.
18. *Rabiner L.R.* A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. Proceedings of the IEEE, 77:257-286, 1989.
19. *Van Rijsbergen, C.J.* Information Retrieval. 2nd edition, Glasgow, University of Glasgow, 1979.