Investigating syllabic prominence with Conditional Random Fields and Latent-Dynamic Conditional Random Fields

Francesco Cutugno[1]*, Enrico Leone*[1]*, Bogdan Ludusan*[2]*, Antonio Origlia*[1]

[1]LUSI-Lab, Dept. of Physics, University of Naples "Federico II", Italy
[2]CNRS-IRISA, Rennes, France

{cutugno, antonio.origlia}@unina.it, erik.leone82@gmail.com, bogdan.ludusan@irisa.fr

Abstract

The present study performs an investigation on several issues concerning the automatic detection of prominences. Its aim is to offer a better understanding of the prominence phenomenon in order to be able to improve existent prominence detection systems. The study is threefold: first, the presence of hidden dynamics in the sequence of prominent and non-prominent syllables is tested by comparing results obtained with CRFs and LDCRFs. Second, the size of the context to be taken into account when determining prominence was examined and third, a new set of features was investigated. The obtained results show that LDCRFs systematically outperform CRFs, that a context of three syllables is generally sufficient for prominence detection and that syllable length is a useful feature to include. Also, new features concerning pitch movements we introduced can substitute adequately heuristic measures used in previous works.
Index Terms: syllabic prominence, conditional random fields

1. Introduction

Linguistic research has concentrated for a long time on the investigation of syllabic prominence. Although there is no consensus regarding the definition of syllabic prominence nor on the appropriate annotation methodology, it is common, especially in automatic annotation studies, to describe prominent syllables in generic terms as *syllables standing out with respect to their context* and to annotate them by means of a binary notation (prominent/non-prominent). This type of annotation is generally preferred because it offers a simple method to evaluate the performance obtained by automatic approaches. Prominence detection has been the subject of a wide number of investigations in the past [1, 2, 3, 4, 5]. Automatic prominence detection systems are based mainly on rule-based approaches as machine learning techniques can make it difficult to understand how a certain performance was reached by the underlying statistical model. Although supervised approaches were used in this work, absolute performance was considered a secondary objective with respect to the possibility of using machine learning to collect data useful to improve current rule-based annotation systems based on a linguistic, as opposed to a statistic, background.

Conditional Random Fields (CRF) [6] are a class of discriminative models used for sequence segmentation and labeling which are designed to maximize the conditional probability of the labels given the sequence of observations. The use of CRFs is now well established as they have been successfully applied to a wide range of scientific fields, including natural language processing and speech analysis tasks. In the case of prominence annotation, it was shown that CRFs outperform HMMs in the task of predicting pitch accents at word level with a combination of acoustical and syntactic features [7]. CRFs were also used to investigate pitch accent detection along with the realization of givenness and focus at word level by employing lexical and acoustic features [8]. Differently from these previous studies, in this work we will make use of acoustic features only and we will concentrate on the syllable level.

There are three main ways in which this particular kind of sequence labeling models can be applied to the problem of automatic syllabic prominence annotation in order to guide further investigation towards a better rule-based annotation algorithm. **Structural differences analysis** among classifiers, paired with performance comparison, gives information regarding the interactions among the features. **Feature sets comparison** and **multiple contexts comparison** estimate the predictive power of the considered features and the amount of context that should be taken into account. In this work, we apply these three different kinds of analysis by employing different classifiers, feature sets and context extensions.

2. Materials

We used an Italian corpus containing read numbers and an English corpus containing read sentences. The Italian corpus consists of a subset of the SPEECON corpus [9] that has been used to evaluate the system presented in [3, 5]. The English corpus consists of a subset of the TIMIT corpus that has been used to evaluate the system presented in [2]. Both subsets were manually segmented into syllables and annotated by an expert linguist using a binary notation for syllabic prominences. The SPEECON subset contains 288 utterances (15 minutes of speaking time) containing at least 5 syllables (mean: 15, total: 4265). The TIMIT subset contains 382 utterances (17 minutes of speaking time) containing at least 4 syllables (mean: 12.51, total: 4780).

3. Latent-Dynamic Conditional Random Fields

CRFs are designed to capture inter-class relationships by maximizing the conditional probability of the sequence of labels from a sequence of observations. Given a set of weights estimated during training λ, the sequence of labels Y and the sequence of observations X, a Linear Chain Conditional Random Field estimates $P(Y|X)$ as follows

$$P(Y|X,\lambda) = \frac{1}{Z(X)} exp\left(\sum_{k=1}^{K} \lambda_k f_k(y_t, y_{t-1}, \mathbf{x}_t)\right) \quad (1)$$

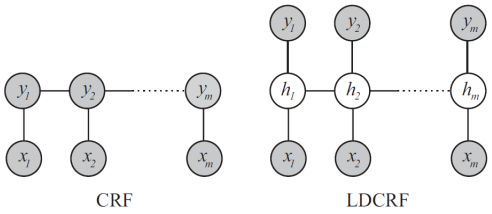

Figure 1: *Graphical representation of a Conditional Random Field and a Latent-Dynamics Conditional Random Field.*

where $Z(X)$ is a normalization constant, N is the number of observations and $f_k(y_t, y_{t-1}, \mathbf{x}_t)$ represents either a *state feature function* or a *transition feature function*. State feature functions describe the relation between observation/label pairs while transition feature functions describe the relation between observations and transitions from one state to another. Since the definition of feature functions includes a vector of observations in the third term, the set of feature functions can be computed over an arbitarily extended context of surrounding observations W. CRFs are limited as they can model inter-class relationships but cannot model intra-class dynamics. Latent Dynamic Conditional Random Fields (LDCRF) [10] are an extension of CRFs designed to introduce hidden variables in the model, in order to capture both kinds of dynamics. Hidden states represent a sequence of unobserved variables H and are used to define the following latent conditional model:

$$P(Y|X,\lambda) = \sum_{H} P(Y|H,X,\lambda)P(H|X,\lambda) \tag{2}$$

The above model allows only disjoint sets of hidden states for each class label. Therefore, each label y_j has an associated set H_{y_j} of hidden states with $H_{y_i} \cap H_{y_j} = \emptyset$ for $i \neq j$, making it is possible to rewrite Equation 2 as:

$$P(Y|X,\lambda) = \sum_{h \in H_{y_j}} P(H|X,\lambda) \tag{3}$$

The conditional probability of the hidden states given the set of observations and weights can then be formulated as for the CRF model:

$$P(H|X,\lambda) = \frac{1}{Z(X)} exp\left(\sum_{k=1}^{K} \lambda_k f_k(h_t, h_{t-1}, \mathbf{x}_t)\right) \tag{4}$$

A graphical comparison between a CRF and an LDCRF is shown in Figure 1. In the LDCRF model there is no longer a direct connection between observations and labels due to the introduction of a layer of hidden variables. Since the labels are disconnected from the observations, they are considered to be conditionally independent, given the hidden states.

Although CRFs have been used in the past to automatically detect prominent syllables, there are no studies, to our knowledge, aimed at evaluating the performance of LDCRFs on the same task.

4. Feature sets

Features related to energy, segments durations and internal pitch movements for each syllable are usually employed in the automatic prominence annotation task and the particularly important role that syllable nuclei play in the detection of prominent syllables is widely recognized in the literature. In [1], the mean amplitude inside the syllable nucleus ΔA and the nucleus length ΔT_n were combined into an evidence variable as follows:

$$Ev = \Delta A \Delta T_n \tag{5}$$

After computing an Ev value for each syllable, local maxima in the sequence of evidence variables were marked as prominent. Since, in the literature concerning automatic annotation of syllabic prominence, a great importance has been assigned to ΔA and ΔT_n, these two features are always included in the feature sets we used in the experiments presented in this paper. Given the manual segmentation into syllables, nuclei onsets and offsets are estimated by taking the energy peak inside the syllable and computing the -3dB band.

Energy and duration do not account for prominences caused by pitch movements through the nucleus. In [4], a syllable was automatically marked as prominent if a rising pitch movement exceeding a threshold was detected. In [3], the same concept was implemented as an integration of the approach proposed in [1] with a pitch movement dependent parameter. Equation 5 was then reformulated as

$$Ev = m\Delta A \Delta T_n \tag{6}$$

where m represented a heuristically computed penalization factor for syllables that do not exhibit a rising movement through the nucleus. The m parameter is included in the feature sets F3 and F4.

The previous two attempts to use pitch features in an automatic system for prominence annotation were based on heuristics and assumed that only rising pitch movements had an effect on prominence perception. In this paper, we introduce a different parameter that is intended to give an account of pitch movements in a more generic term, by using the concept of *glissando*, or dynamic tone. Let T_e be the time interval in which a pitch movement is realized and the rate of change expressed in ST/s, the threshold over which a tone is perceived as dynamic instead of static was estimated in [11] to be $0.16/T_e^2$. Later on Mertens, in [12], found that an automatic pitch curve stylization algorithm based on a tonal perception model, taking glissandos into account, gave results more similar to the ones provided by human annotators by setting the glissando threshold at $0.32/T_e^2$. In this paper, we introduce a new feature related to the occurrence of glissandos through the nucleus that is intended to substitute the m parameter. After stylizing the pitch curve using the basic version of the algorithm presented in [13], we consider the position of the energy peak inside the manually marked syllable and select the linear segment that crosses it. The two extrema of the segment, s_1 and s_2, are then taken as reference to compute the rate of change of the movement crossing the nucleus. If the s_1 and s_2 points fall outside the syllable nucleus, their position is moved respectively to the point where the segment crosses the nucleus onset or offset. A Γ_{s_1,s_2} parameter is then computed as follows

$$\Gamma_{s_1,s_2} = \begin{cases} 1 & if\ V_{s_1,s_2} > \frac{0.32}{T_e^2} \\ \frac{V_{s_1,s_2} T_e^2}{0.32} & otherwise \end{cases} \tag{7}$$

where V_{s_1,s_2} is the absolute rate of change of the pitch excursion in ST/s. While a tonal movement exceeding Mertens' threshold will be given 1 as value, movements below the threshold will be described with a glissando likelihood value comprised between 0 and 1. The Γ_{s_1,s_2} parameter has a stronger

theoretical background with respect to the m parameter and is to be preferred because it avoids assumptions regarding the magnitude of pitch change that is needed to introduce a prominence, considering only the presence of a glissando. The Γ_{s_1,s_2} parameter reacts both to rising and descending movements and is based on results obtained by previous investigations regarding dynamic tones perception.

We also evaluate the impact of two other features extracted from each syllable: the total duration ΔT_s and the internal ratio between voiced and total duration of the syllable V/T_s. ΔT_s was included as it is common, in the literature, to find documentation regarding the importance of duration features in prosodic analysis while V/T_s is intended to give an account of the considered syllable structure. These two features are included in the features set F5. A detailed description of the composition of the three feature sets tested in this work is shown in Table 1.

Table 1: *Feature sets composition*

	ΔA	ΔT_n	m	Γ_{s_1,s_2}	ΔT_s	V/T_s
F3	✓	✓	✓			
F4	✓	✓	✓		✓	
F5	✓	✓		✓	✓	✓

5. Results

Each classifier was tested on the SPEECON and on the TIMIT subset. For each features set, both classifiers were tested by varying the context extension for building the feature functions from a minimum of 1, which considers only the two neighboring syllables, to a maximum of 5. Performance is measured in terms of F-measure (Prominent class as TRUE) and the test protocol is 10-fold cross validation. The summary of the results obtained on the SPEECON subset is reported in Table 2 while results obtained on the TIMIT subset are detailed in Table 3.

Table 2: *F-measures obtained on the SPEECON subsets. The best performance obtained with each features set by the two classifiers is marked in bold.*

	CRF			LDCRF		
	F3	F4	F5	F3	F4	F5
W1	65.12	79.79	82.76	65.09	79.69	82.75
W2	67.10	82.08	84.30	75.10	84.66	86.32
W3	68.00	83.46	85.77	75.39	85.58	87.82
W4	68.27	84.05	**86.15**	76.12	**85.71**	87.66
W5	**68.94**	**84.07**	86.08	**76.55**	85.54	**87.85**

Table 3: *F-measures obtained on the TIMIT subset. The best performance obtained with each features set by the two classifiers is marked in bold.*

	CRF			LDCRF		
	F3	F4	F5	F3	F4	F5
W1	53.43	68.50	68.91	53.52	68.56	68.90
W2	60.47	71.71	70.96	72.03	**78.01**	77.24
W3	60.43	71.91	71.54	**72.52**	77.83	**77.52**
W4	61.46	72.07	72.03	72.12	77.86	77.26
W5	**61.76**	**72.36**	**72.48**	72.12	77.83	77.21

The statistical significance of the differences between the obtained performances was evaluated by means of a McNemar test. To evaluate the performance of the LDCRF with respect to the CRF, we compared, for each features set, the results obtained by the best CRF with the ones obtained by the best LDCRF. The differences were found to be statistically significant in all cases ($p < 0.01$).

To evaluate the performance difference obtained with the various feature sets, we compared the performance of the best LDCRF from each features set with the performance of the best LDCRFs from the other feature sets. While on the SPEECON subset the differences were found to be always significant ($p < 0.001$), on the TIMIT subset the differences F3/F4 and F3/F5 were statistically significant ($p < 0.001$) while the difference F4/F5 was not significant.

To evaluate the influence of the context extension, we compared, for each corpus, classifier and features set, the pairwise combinations of values of the W parameter. While on the TIMIT subset only comparisons involving $W = 1$ were found to be significant, tests on the SPEECON subset yielded a different situation, summarized in Table 4.

Table 4: *Statistical significance tests on the SPEECON corpus, for different pairs of values for the W parameter. Check marks indicate significant differences.*

		Window length pairs					
		2/3	2/4	2/5	3/4	3/5	4/5
CRF	F3						
	F4	✓	✓	✓			
	F5	✓	✓	✓			
LDCRF	F3				✓	✓	
	F4				✓		
	F5	✓	✓	✓			

Both LDCRFs and CRFs applied to the SPEECON subset outperform the systems presented in [3], in which 73.3% F-measure was reported, and in [5], where 75.1% F-measure was reported. Concerning the TIMIT corpus, in [2] an error rate of 18.64% was reported. If we take into account the performance of the LDCRF that obtained the best results on TIMIT (F4/W2), the error rate is 16.32%. A graphical summary of the presented tests is shown in Figure 2.

6. Discussion

Results offer insight on three different issues regarding prominence detection: model performance, context influence and feature sets. We found that LDCRFs perform systematically better than CRFs. On the SPEECON subset, every LDCRF performs better than its direct CRF counterpart while on TIMIT this effect is even more evident as the lowest performance obtained by an LDCRF is similar to the best CRF performance. The main difference between the two classifiers lies in the presence, in the LDCRF model, of hidden states. This difference is critical as it allows the classifier to learn complex dynamics that are not explicitly described by the raw sequence of observations. In order to better understand these results two observations are important, in our opinion: (1) the advantage of LDCRFs is that they detect hidden dynamics inside a single class and (2) the binary annotation mainly produces sequences of non-prominent syllables separated by prominent syllables. A possible cause of LDCRFs outperforming CRFs in this task could, therefore, be that a hidden dynamic may lie in the sequence of non-prominent syllables.

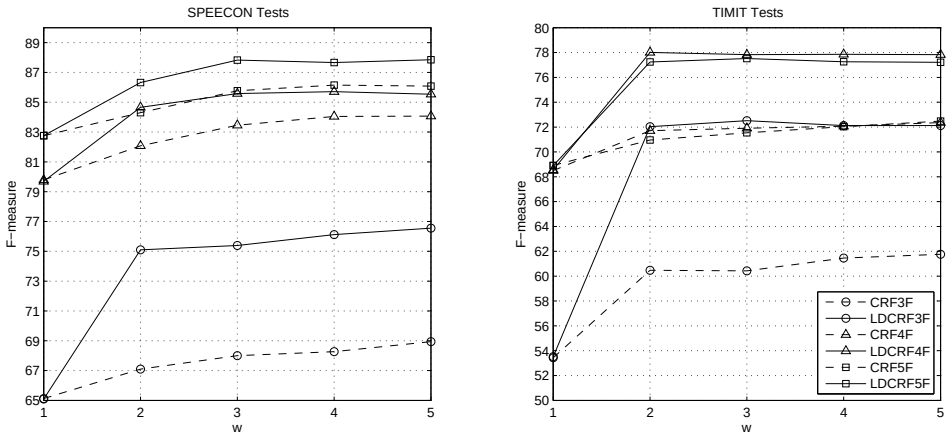

Figure 2: *Summary of the obtained performances for each combination of classifier, features set and context extension on the two test corpora.*

By varying the context extension, we observed that on the SPEECON subset significant differences among various tests can be found consistently up to a three syllables context. On the TIMIT subset a context window of two syllables seems to be sufficient to achieve maximum performance. This is in contrast with approaches used in earlier automatic prominence annotation [1], but it is consistent with more recent findings regarding context extension [4].

By varying the composition of the features set, we found that syllable length is a particularly important feature as the comparison between the best LDCRF using the F3 features set is always inferior to the performance obtained by the best LDCRF using the F4 features set in a statistically significant way. Since the F5 features set enabled the LDCRF to obtain better performance with respect to the F4 features set on the SPEECON subset only, we understand that the combination of the $\Gamma_{s1,s2}$ and V/T_s features contains at least the same amount of information as the m parameter, while having a better theoretical background.

7. Conclusions

We investigated the problem of automatic syllabic prominence detection using two classifiers designed to label sequences of data: the CRF and the LDCRF. Other than the comparison between the two classifiers, we studied the differences between three different feature sets and the impact that context extension has on the results. The systematically higher performance of LDCRFs with respect to CRFs suggests the presence of hidden dynamics in the analyzed sequences that may influence prominence distribution over the utterance. From our experiments on context extension, we found indications that a window of 2-3 neighboring syllables contains the necessary amount of data that allows an automatic annotation algorithm to obtain good performance in terms of F-measure. Lastly, we compared the performance of the best classifiers with different feature sets observing a significant importance of syllable length. We also observed that the combination of two new features substituting heuristic approaches to pitch movements description improved the performance of the classifier on the SPEECON subset and did not lower the performance on the TIMIT subset.

8. Acknowledgements

The authors appear in strict alphabetic order. We would like to thank Hugues Salamin from the University of Glasgow for his useful insight for the use of CRFs and LDCRFs.

9. References

[1] R. Silipo and S. Greenberg, "Automatic transcription of prosodic stress for spontaneous English discourse," in *Proc. of ICPhS*, 1999, pp. 2351–2354.

[2] F. Tamburini, "Reliable prominence identification in english spontaneous speech," in *Proc. of Speech Prosody [Online]*, 2006.

[3] G. Abete, C. Cutugno, B. Ludusan, and A. Origlia, "Pitch behavior detection for automatic prominence recognition," in *Proc. of Speech Prosody [Online]*, 2010.

[4] M. Avanzi, A. Lacheret-Dujour, and B. Victorri, "A corpus based learning method for prominence detection inspontaneous speech," in *Proc. of Speech Prosody [Online]*, 2010.

[5] B. Ludusan, A. Origlia, and F. Cutugno, "On the use of the rhythmogram for automatic syllabic prominence annotation," in *Proc. of Interspeech*, 2011, pp. 2413–2416.

[6] J. D. Lafferty, A. McCallum, and F. C. N. Pereira, "Conditional random fields: Probabilistic models for segmenting and labeling sequence data," in *Proc. of ICML*, 2001, pp. 282–289.

[7] M. L. Gregory, "Using conditional random fields to predict pitch accents in conversational speech," in *Proc. of ACL [Online]*, 2004.

[8] V. K. R. Sridhar, A. Nenkova, S. Narayanan, and D. Jurafsky, "Detecting prominence in conversational speech: pitch accent, givenness and focus," in *Proc. of Speech Prosody [Online]*, 2008, pp. 453–456.

[9] R. Siemund, H. Hge, S. Kunzmann, and K. Marasek, "Speecon-speech data for consumer devices," in *Proc. of LREC*, 2000, pp. 883–886.

[10] L. Morency, A. Quattoni, and T. Darrell, "Latent-dynamic discriminative models for continuous gesture recognition," in *Proc. of CVPR*, 2007, pp. 1–8.

[11] J. t'Hart, R. Collier, and A. Cohen, *A Perceptual Study of Intonation: An Experimental-Phonetic Approach*. Cambridge: Cambridge University Press, 1990.

[12] P. Mertens, "The prosogram: Semi-automatic transcription of prosody based on a tonal perception model," in *Proc. of Speech Prosody [Online]*, 2004.

[13] A. Origlia, G. Abete, C. Cutugno, I. Alfano, R. Savy, and B. Ludusan, "A divide et impera algorithm for optimal pitch stylization," in *Proc. of Interspeech*, 2011, pp. 1993–1996.