An Evaluation of Machine Learning Methods for Prominence Detection in French

George Christodoulides [1], *Mathieu Avanzi* [2]

[1] Centre Valibel, Institute for Language & Communication, Université de Louvain, Belgium
[2] UMR 7110-LLF (Laboratoire de Linguistique Formelle), Université Paris-Diderot

george@mycontent.gr, mathieu.avanzi@gmail.com

Abstract

The automatic detection of prosodically prominent syllables is crucial for analysing speech, especially in French where prominence contributes substantially to prosodic grouping and boundary demarcation. In this paper, we compare different machine learning techniques for the automatic detection of prominent syllables, using prosodic features (including pitch, energy, duration and spectral balance) and lexical information. We explore the differences between modelling the detection of prominent syllables as a classification or as a sequence labelling problem, and combinations of the two techniques. We train and evaluate our systems on a corpus of spontaneous French speech, consisting of almost 100 different speakers; the corpus is balanced for speaker age and sex and covers 3 different regional varieties. The result of this study is a novel tool for the automatic annotation of prominent syllables in French.

Index Terms: prosodic prominence, automatic detection, French

1. Introduction

Modelling prosodic prominence is of crucial importance for the speech research community, as demonstrated by the numerous publications on this topic during the past decade. There is a general agreement in the literature on the following observations: (i) prominence is defined as the phenomenon of a linguistic unit (syllable, word, or even larger stretch of speech) being perceived as standing out of its environment ([1], [2]); (ii) the perception of prominence is influenced by various constraints, ranking from low-level acoustic cues to top-down expectations (e.g. [3], [4], [5]); (iii) prosodic prominence can be automatically detected by using a reduced set of acoustic parameters (including pitch, duration, energy and spectral features) and lexical information (part-of-speech tags). In French, syllabic prominence is of crucial importance because it essentially contributes to mark the boundaries of prosodic groups (unlike variable-stress languages, such as English or Dutch, where 'stress' is a lexical property of a specific syllable in a word, while 'accents' are used to signal the information status of a linguistic unit).

Experiments on the perception of prosodic prominence have shown that the reliability between two or more annotators can reach a satisfactory level (e.g. [6], [7], [8]). It is thus estimated that the accuracy between trained and experimented human raters lies between 85% and 90% at best ([9], [10], [11]). At the same time, efforts have been made to build robust algorithms for automatic prominence detection. When compared with human annotation, the accuracy of unsupervised methods for prominence detection barely exceeds 80% (e.g. [12], [13]); supervised methods –i.e. algorithms trained on transcribed and labelled material– may reach an accuracy of 80% to 90% (accuracy being the ratio of units correctly labelled to the total number of units). For the task of detecting syllabic prosodic prominence, [14] use a corpus-based learning method (inferring rules from an analysis of a labelled corpus) and obtain an F-measure of almost 80%. Using the Random Forest classification method, [11] presents a model that explains 85% of the observed variance. An approach based on Conditional Random Field (CRF) and Latent Dynamic CRFs (LDCRF) is presented in [15], reporting an F-measure of 73.3% and 75.1% respectively. Using various classification models and linear discriminant analysis (LDA) to select the most promising features, [16] reach an F-measure between 84.1% and 87.5%.

These datasets on which the above-mentioned algorithms are evaluated and/or trained include different speaking styles (read speech and spontaneous speech, for most studies), and up to 30 speakers. The corpus size is measured in syllables, ranging between 6k and 17k syllables; they correspond to corpora between 20 min and 120 min-long.

We present a study on a significantly larger corpus, annotated by experts for syllabic prosodic prominence. We evaluate several different algorithms for automatic prominence detection, framing the task either as a classification problem, or as a sequence labelling problem. We evaluate the performance of these algorithms and propose a new tool for the automatic detection of syllabic prominence in spoken French.

2. Data

2.1 Material and participants

The corpus used in this study consists of recordings extracted from the PFC database ([17], [18]). It includes speech material recorded in 14 geographical areas, spread over 3 European French-speaking countries: 5 varieties spoken in Metropolitan France (Béthune, Brécey, Lyon, Paris and Ogéviller); 5 varieties spoken in Switzerland (Fribourg, Geneva, Martigny, Neuchâtel and Nyon) and 4 varieties spoken in Belgium (Brussels, Gembloux, Liège and Tournai). For each of the 14 sites, 4 female and 4 male speakers were selected; they were born and raised in the city in which they were recorded. The age of the speakers varies between 20 and 80 years. The corpus is stratified into four age groups: this parameter is controlled for each of the 14 groups of speakers ($F(13, 84) = 0.308$), between male and female speakers ($F(1, 84) = 0.110$, n.s.) and between male and female speakers across the 14 groups ($F(13, 84) = 0.114$, n.s.).

Each speaker was recorded in a reading text task (the text is 398 words-long) and in semi-directed sociolinguistic interviews, in which the informant has minimal interaction with the interviewer. The entire reading text and a stretch of 3 minutes of spontaneous speech for each speaker were orthographically transcribed and automatically aligned within

Praat [19] with the *EasyAlign* script [20], which provides a 3-layer annotation in phones, syllables and words. All alignments were manually verified and corrected when necessary by inspecting both spectrogram and waveforms. All the alignments were checked and corrected by hand, by inspecting waveforms and spectrograms. The orthographic transcription was then annotated with part-of-speech tags using the *DisMo* software [21]. A dedicated tier ("delivery") was added, in order to annotate overlapping speech and short non-audible or unusable segments (e.g. due to the presence of external noises in the recording). In total, the corpus is approximately 11 hours-long, and includes approximately 113 thousand tokens (63k in semi-directed interviews and 47k in reading). Table 1 presents the basic properties of this corpus.

Table 1. *Corpus used for training and evaluation.*

Area	Region	Age		Dur. (sec)	Nb. syll.	Nb. tokens
		Min-Max	Mean (sd)			
Belgium	Brussels	27-65	44 (15)	2810	11446	8565
	Gembloux	22-76	42 (21)	2821	11677	9135
	Liège	21-76	48 (24)	2951	9692	7400
	Tournai	19-82	44 (26)	2837	10518	8031
France	Béthune	21-89	46 (25)	2925	11153	8571
	Brécey	19-80	47 (22)	3110	11505	8659
	Lyon	21-74	42 (21)	2677	10866	7783
	Paris	24-86	50 (22)	2896	10088	8188
	Ogéviller	23-93	58 (24)	3023	10685	8101
Switzerland	Fribourg	20-82	43 (24)	2895	10865	8186
	Geneva	21-61	41 (18)	2863	10720	8062
	Martigny	22-80	49 (28)	2963	10199	7726
	Neuchâtel	25-78	53 (24)	2960	10201	7625
	Nyon	30-70	46 (17)	2929	9948	7637
Totals	**14 points**	**19-93**	**46 (21)**	**11.2 h**	**149 563**	**113 669**

2.2 Labelling

Prominent syllables and syllables associated with a disfluency (fillers, lengthened syllables due to hesitations, false starts, repairs, etc.) were identified independently by two experts on the basis of their perceptual judgment only, following the C-PROM methodology, presented in detail in [10]. A third expert intervened in cases of disagreement between the two annotators and decided the final value of the syllable (+/- prominent, +/- associated to a disfluency) in a dedicated tier [18]. Data labelling was performed over a period of almost three years. Four couples of annotators (every team including the second author of this paper) took turn. Kappa statistics [22] were used to assess the reliability for each pair regarding prominence annotation, and lead to Kappa values varying between 0.61 and 0.88, with a mean of 0.72, which is considered as fair and perfect according to [23].

3. Method

The dataset was processed by using *Praaline* [24], a toolkit that interfaces with *Praat* and runs a cascade of scripts and/or external analysis tools, each of which may add features to an annotation level, stored in a relational database. Using this interface, we applied *Prosogram* [25] for pitch stylisation on the entire corpus. *Prosogram*'s algorithm operates in two phases; for each syllable, vocalic nuclei are detected based on intensity and voicing. The f_0 curve on the nucleus is then stylised into a static or dynamic tone, based on a perceptual glissando approach. After this pre-processing step, a number of acoustic, prosodic and lexical features were extracted for each syllable, including:

- Syllable duration (ms);
- Minimum, maximum and mean pitch (f_0);
- Pitch movement (within syllable and between successive syllables);
- Peak intensity within the syllabic nucleus;
- Spectral balance;
- Token (word) to which the syllable belongs, and Part-of-Speech tag of the token;
- Presence and duration of subsequent pause;
- Syllabic structure (C/V, whether the syllable ends with a schwa);
- Position of the syllable relative to the token (word) in which it belongs: initial, final, penultimate, mono-syllabic word.

The selection of the above-mentioned features is based on the findings in previous research relating to the perception of syllabic prominence in spoken French. Relative measures were calculated for each syllable, with different contexts: the mean value of a measure is calculated over a symmetric window of 2, 3, 4 and 5 syllables (before and after the current one); silent pauses block the context window. Pitch and intensity measures are in logarithmic scales (pitch is converted to semitones relative to 1 Hz; intensity is in dB). A z-score transformation was applied to normalize the data over each corpus recording (sample), thus taking into account speaker variation (e.g. the differences in pitch register between different speakers). We followed the definition of spectral balance outlined in [26]: according to this approach, the mean energy of four adjacent frequency bands is measured (0-500 Hz, 0.5-1 kHz, 1-2 kHz and 2-4 kHz). All calculated features were stored in the database.

Several machine learning algorithms were tested, for the task of predicting whether a syllable should be labelled as prominent or non-prominent. Despite the fact that previous research has shown that prominence is perceived as a gradual phenomenon (see among other [5], [27]), we frame the problem as a binary classification question, as (i) the experts in charge of the data labelling were asked to annotate syllables as prominent or not, (ii) this is the most practical approach for an automated system whose outputs will be used in later stages of processing (e.g. for minor prosodic phrasing annotation).

This task is traditionally treated as a classification problem: we thus tested a Decision Tree classification algorithm (attribute selection performed based on the Information Gain criterion) [28], Support Vector Machines (C-SVM with an RBF kernel), Neural Network classification (ANN with 20 hidden layer neurons), and a Random Forest (30 decision trees) [29]. However, the annotation of prosodic prominence can also be defined as a sequence labelling problem: given a sequence of syllables, separated by silent pauses, a machine learning method attributes a prominent / non-prominent label to each syllable, based on its features and its context. A particularly suitable method for this type of annotation is based on Conditional Random Fields (CRF) [30].

We conducted an initial testing of the performance of classification algorithms using a reduced set of features (see infra) using the *Orange* [31] data mining framework[1]. The most promising classifiers (Random Forest and SVM), as well as the CRF sequence labelling method, were selected for further testing. To fine-tune the parameters of each algorithm and perform detailed cross-validation tests, we developed a plugin for *Praaline*, in C++ using the open source libraries RF-ACE [2], libSVM [3] and CRF++ [4]. This tool allows for the selection of different sets of features, training models on different subsets of the dataset and testing the prediction accuracy, precision, recall and F-score[5].

4. Results

4.1 Performance of different methods

The results of the preliminary testing of algorithms are shown in Table 2. A sub-corpus of 49k syllables and a reduced set of features (relative duration, relative max pitch, relative mean pitch, relative intensity within a context of 2, 3 and 4 syllables; presence of pause; position of syllable within word) were used for training. The evaluation was performed using 5-fold cross validation.

Table 2. *Performance of classification algorithms on 49k syllables (19 features; see text) using 5-fold cross-validation.*

Algorithm	Accuracy	Precision	Recall	F1 measure
Random Forest	83.3%	86.3%	69.2%	76.8%
SVM	83.1%	82.8%	72.8%	77.5%
Neural Network	84.2%	83.9%	74.8%	79.1%
Classification Tree	77.0%	71.0%	71.6%	71.3%

We compared more robust versions of the Random Forest and SVM classifiers (with the full feature set and after model tuning) with a CRF model. For each feature in position i, its context attributes include $i-2$, $i-1$, i, $i+1$, $i+2$, $i \mid i-1$, $i \mid i+1$. In order to train the CRF model it was necessary to discretize the numerical features: we used 10 equally-spaced bins (note that acoustic features such as pitch are already log-transformed; a z-transformation is applied to the relative syllable duration attributes).

Additional features were included in this second set of experiments. In the CRF model, we added lexical attributes, such as the part-of-speech tag (category, such as adjective or pronoun; and full tag), and the word itself. Hesitations and disfluencies are not excluded from the CRF sequence (they are however marked; thus the model never assigns syllabic prominence annotation to a filled pause). For the SVM classifier, we converted categorical attributes to a vector of binary attributes (e.g. {A, B, C} is represented by A = (1, 0, 0), B = (0, 1, 0), C = (0, 0, 1)). We also evaluated the performance of a CRF model that included the result of the RF classifier as an additional attribute. The results are presented in Table 3.

[1] http://orange.biolab.si/
[2] http://code.google.com/p/rf-ace/
[3] http://www.csie.ntu.edu.tw/~cjlin/libsvm/
[4] http://crfpp.googlecode.com/svn/trunk/doc/index.html
[5] http://www.corpusannotation.org/promise

Table 3. *Performance of different systems for the automatic detection of prominence (full set of features, 5-fold cross-validation)*

Algorithm	Precision	Recall	F1 measure
Random Forest	86.3%	70.4%	77.5%
SVM	83.2%	74.1%	78.4%
CRF	83.7%	82.5%	83.1%
CRF + RF	85.4%	84.4%	84.9%

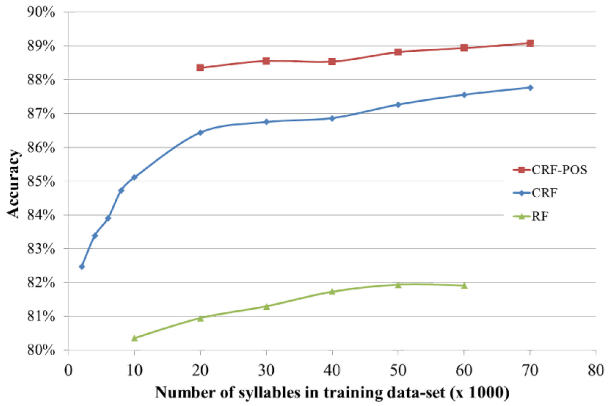
We observe that the CRF model offers an improved F-measure compared to the classifiers. Since series of several non-prominent syllables alternate with prominent ones, the model's ability to capture context relationships is useful. It has to be noted that context relationships are represented in two ways in this model: by the relative features (calculated over different context windows), and by the CRF context attributes themselves.

It may not be practical, however, to include lexical and part-of-speech information in an automatic annotator, since the detection of prosodic prominence may precede or be performed independently of POS tagging. Therefore, in the next set of experiments, we calculated the improvement in accuracy gained by including such information in a discrete CRF model.

4.2 Effects of the training dataset size

In a third set of experiments, we compared the accuracy (percentage of correctly labelled syllables over all syllables) of three algorithms, for different sizes of the training data-set, and with / without lexical and POS information. Figure 1 shows the evolution of the system's accuracy as the training data-set grows from 1,000 syllables to 60,000 syllables. The algorithms compared are: Random Forest based only on acoustic features; discrete CRF based only on acoustic features; and discrete CRF based on acoustic features and the part-of-speech tags (of tokens corresponding to each syllable).

Figure 1. *Effects of the training dataset size on the accuracy of different methods.*



We note that the CRF models were systematically more accurate than the RF method, when based on the same acoustic features. The CRF models seem to stabilize at a corpus size of approximately 20k syllables; beyond this point, accuracy does improved by training on larger datasets, but at a much slower rate. Adding access to the lexical and part-of-speech information to the CRF model increases its accuracy by 2% (cf. [32] and [33] on the influence of part-of-speech categories on the perception of prominence).

5. Conclusions

In this study we have presented the results of training and evaluating different machine learning algorithms for the detection and annotation of syllabic prosodic prominence in spoken French corpora. Among the classification techniques tested, the Random Forest-based ensemble learning gave best results. Overall, however, the CRF-based sequence labelling approach gave the best results; the combination of classifiers with CRF labellers marginally improved performance. Including lexical and part-of-speech information may be used to model the top-down expectations of native listeners with regards to prominence; however, it is possible to create a robust detection system based only on acoustic features. Our experiments suggest that a direction for future studies is finding a more comprehensive set of acoustic correlates.

To the best of our knowledge this is the first study to have used an 11-hour-long corpus, with 2 different speaking styles, to train a system for the automatic detection of prominences. In future studies we plan to explore regional differences in prosodic prominence, as well as the relationship between the expert annotation and naïve perception.

Within the context of this study, a simplified tool to annotate syllabic prominence in French spoken corpora was developed: the tool is called ***Promise***, and is available under an open-source GPL3 license. It works either as a plugin for *Praaline* (and thus may be included in a corpus processing workflow), or as a stand-alone tool with a basic interface. The tool may be downloaded from:

www.corpusannotation.org/promise

6. References

[1] Terken, J., "Fundamental Frequency and Perceived Prominence", Journal of the Acoustical Society of America, vol. 89, pp. 1768–1776, 1991.

[2] Terken, J. et Hermes, D. "The perception of prosodic prominence", in M. Horne (Ed.) Prosody: Theory and experiment. Studies presented to Gosta Bruce, Dordrecht: Kluwer, pp. 89–127, 2000.

[3] Arnold, D. and Wagner, P., "The influence of top-down expectations on the perception of syllable prominence", in ISCA Workshop on Experimental Linguistics, pp. 25-28, 2008.

[4] Cole, J., Mo, J. and Hasegawa-Johnson, M., "Signal-based and expectation-based factors in the perception of prosodic prominence", Laboratory Phonology, vol. 1, pp. 425–452, 2010.

[5] Wagner, P., Tamburini, F., and Windmann, A., "Objective, Subjective and Linguistic Roads to Perceptual Prominence. How are they compared and why?" Proc. of Interspeech, 2012.

[6] Buhmann, J., Caspers, J., van Heuven, V., Hoekstra, H., Martens, J-P., and Swerts, M., "Annotation of Prominent Words, Prosodic Boundaries and Segmental Lengthening by Non Expert Transcribers in the Spoken Dutch Corpus", Proc. of LREC, pp. 779–785, 2002.

[7] Mo, Y., Cole, J. and Lee, E., "Naive Listeners' Prominence and Boundary Perception", Proc. of Speech Prosody, 735-738, 2008.

[8] Smith, C. "Perception of prominence and boundaries by naïve French listeners", Proc. of the 17th ICPHS, 1874–1877, 2011.

[9] Wightman, C., and Ostendorf, M., "Automatic labeling of prosodic patterns", IEEE Transactions on Speech and Audio Processing, vol. 2, no. 4, pp. 469–481, 1994.

[10] Avanzi, M., Simon, A.C., Goldman, J-Ph., and Auchlin, A., "C-PROM: An Annotated Corpus for French Prominence Study", Proc. of Prosodic Prominence, Speech Prosody Workshop, 2010.

[11] Arnold, D., Wagner, P., and Baayen, H., "Using generalized additive models and random forests to model German prosodic prominence", Proc. Interspeech, 272–276, 2013.

[12] Tamburini, F., "Automatic Prosodic Prominence Detection in Speech using Acoustic Features: an Unsupervised System", Proc. of Interspeech, 129–132, 2003.

[13] Goldman, J-Ph., Avanzi, M., Simon, A.C., Lacheret-Dujour, A., and Auchlin, A., "A Methodology for the Automatic Detection of Perceived Prominent Syllables in Spoken French", Proc. of Interspeech, 98–101, 2007.

[14] Avanzi, M., Lacheret-Dujour, A., and Victorri, B., "A Corpus-based Learning Method for Prominence Detection in Spontaneous Speech", Proc. of Prosodic Prominence, Speech Prosody Workshop, 2010.

[15] Cutugno, F., Leone, E., Ludusan, B., and Origlia, A., "Investigating Syllabic Prominence With Conditional Random Fields and Latent-Dynamic Conditional Random Fields", Proc. of Interspeech, 2402–2405, 2012.

[16] Obin, N., Rodet, X., and Lacheret-Dujour, A., "A Syllable-Based Prominence Detection Model Based on Discriminant Analysis and Context-Dependency", Proc. of Speech and Computer, 2009.

[17] Durand, J., Laks, B., and Lyche, C. [Eds], Phonologie, variation et accents du français. Paris: Hermes, 2009.

[18] Avanzi, M., A Corpus-Based Approach to French Regional Prosodic Variation, Proc. 3rd Swiss workshop on Prosody, Geneva, 2014.

[19] Boersma, P. and Weenink, D. "Praat: doing phonetics by computer (Version 5.5)". www.praat.org, 2014.

[20] Goldman, J.-Ph. "EasyAlign: an automatic phonetic alignment tool under Praat", proc. of Interspeech, pp. 3233-3236, 2011.

[21] Christodoulides, G., Avanzi, M. and Goldman, J.-P. "DisMo: A Morphosyntactic, Disfluency and Multi-Word Unit Annotator. An Evaluation on a Corpus of French Spontaneous and Read Speech", Proc. of LREC, 2014.

[22] Cohen, J. "A Coefficient of Agreement for Nominal Scales", Educational and Psychological Measurement, 20(1):37-46, 1969.

[23] Landis, J. R. and Koch, G. "The Measurement of Observer Agreement for Categorical Data". Biometrics, 33:159-174, 1977.

[24] Christodoulides, G., "Praaline: integrating tools for speech corpus research", Proc. of LREC, 2014.

[25] Mertens, P., "The Prosogram: Semi-automatic transcription of prosody based on a tonal perception model" in B. Bel & I. Marlien [Eds], Proceedings of Speech Prosody 2004, Nara, Japan, 23-26 March, 549-552, 2004.

[26] Sluijter, A. M. C.; van Heuven, V.J. "Spectral balance as an acoustic correlate of linguistic stress", Journal of the Acoustical Society of America 100(4), pp. 2471-2485, 1996.

[27] Goldman, J-Ph., Avanzi, M., Simon, A.C., and Auchlin, A., "A Continuous Prominence Score Based on Acoustic Features", Proc. of Interspeech, pp. 2454-2457, 2012.

[28] Quinlan, J. R., "Induction of Decision Trees", Machine Learning 1: 81-106, Kluwer Academic Publishers, 1986.

[29] Breiman, L., "Random Forests", Machine Learning, 45(1): 5–32, 2001.

[30] Lafferty, J., McCallum, A., and Pereira, F., "Conditional random fields: Probabilistic models for segmenting and labeling sequence data", Proc. of the 18th International Conference on Machine Learning, Morgan Kaufmann, 282–289, 2001.

[31] Demšar, J., Curk, T., and Erjavec, A. "Orange: Data Mining Toolbox in Python", Journal of Machine Learning Research 14(Aug): 2349−2353, 2013.

[32] Goldman, J.-Ph., Auchlin, A., Roekhaut, S., Simon, A-.C. and Avanzi, M., "Prominence perception and accent detection in French. A corpus-based account", Proceedings Speech Prosody 2010, Chicago, 11-14 May, 2010.

[33] Lacheret, A., Simon, A.-C., Goldman, J-Ph. and Avanzi, M., "Prominence perception and accent detection in French: from phonetic processing to grammatical analysis", Language Sciences, vol. 39, pp. 95-106, 2013.