A Syllable-Based Prominence Detection Model Based on Discriminant Analysis and Context-Dependency

Nicolas Obin, Xavier Rodet, Anne Lacheret-Dujour

A Syllable-Based Prominence Detection Model
Based on Discriminant Analysis and Context-Dependency

Nicolas Obin, Xavier Rodet

IRCAM
Analysis-Synthesis team
1, place Stravinsky
75004 Paris, France

Anne Lacheret-Dujour

Université Paris X, MoDyCo lab.
92001 Nanterre, France
& Institut Universitaire de France
75005 Paris, France.

Abstract

On the basis of our previous work, we propose a syllable-based prominence detection model within the framework of exploratory data analysis and discriminant learning in the acoustic domain. This paper investigates two hypothesis on the acoustic data processing: a linear discriminant analysis in which the relative discriminant ability of single prosodic cues are combined into prosodic patterns and a context-dependant model that accounts for phonological dependencies (phonetic intrinsic properties and coarticulation effect). The proposed approach significantly outperforms a baseline method on a corpus of French read speech with a performance of 87.5% in f-measure for the prominent syllables (respectively 90.4% in global accuracy).

1. Introduction

To enable the development of linguistic study and prosodic models over large corpora, a number of tasks should be automatically preprocessed such as phonetic decoding and segmentation, detection of para-verbal phenomena (fillers), prosodic saliencies (prominences) and disfluencies. Research on prominence (syllable or word based) has received a significant amount of attention over past years both in the study of its perception, its acoustic correlates as its automatic detection. We will focuse here mostly on this acoustic dimension.

The perception of prominence is complex: it results from the interaction between acoustic perceptual salience (a bottom-up process involving significant acoustic deviation from more or less abstract references) and cognitive representations and associated expectations (a top-down process which involves lexical feature, syntactical structure, and informational structure)[1, 2]. Most of the studies on prominence, if they do not deny its top-down dimension, have nevertheless focused on its acoustic correlates [3, 4, 5, 6, 7, 8, 9, 10]. These studies have pointed out that the perception of prominence is based on a set of acoustic correlates that seems to be independent of language (variations in duration, pitch [3, 7], intensity, and vocal quality (spectral emphasis, vocal effort) [6, 5, 9, 10]), but whose relative contributions is language-specific [4]. Another property is the acoustical context-dependencies which account for the local acoustic reference baseline used by listeners to judge perceptual saliency [3, 4, 7].

Such findings, while they are fundamental to understand the mechanisms that underlay the perception of prominence in speech, could be of little interest in an automatic detection task. There are two reasons for this: 1) a significant difference observed between acoustic means does not ensure a good separability for each particular observation. 2) the use of controlled utterances neutralizes the phonetic contextual factor which is a source of variability in the acoustic observations (acoustic properties intrinsic units observed and coarticulation) in speech. These two reasons combined together explain why a relevant acoustic cue in the case of laboratory speech may have a low discriminating ability or difficult to integrate in a prominence automatic detection task [11].

In parallel, methods have been proposed for the automatic detection of prominence [12, 13, 14, 15, 16, 17, 11, 18]. Research in the field of automatic detection of prominence has mainly concentrated on the search and integration of discriminant acoustic features [13, 15, 11, 18], relatively few on the choice of classifiers. On the acoustic level: pitch (f_0) and duration feature are known and commonly used. If these correlates show robust performance in a prominence detection task, it seems that they have limited discriminant ability and are not sufficient in a significant number of cases. Therefore, more recent studies have tried to investigate some new acoustic correlates such as energy (acoustic and perceived)[15], energy bands [11], and spectral emphasis. These methods are based on the discriminant ability of a set of single acoustic cues; [13] has proposed to combine a set of acoustic cues in order to cumulate their relative contributions in the achievement of an acoustic salience. However, the proposed method is based on some heuristics postulated solely on the basis of expert knowledge. On the classifier level: if the decision tree is referred as a baseline classifier, some other classifiers have been proposed through literature (neural networks, HMM, rule-based [12], bagging and boosting approach [14], coupled-HMM[16], voting classifiers [11]).

Following our previous work [18], our approach is based on exploratory data analysis and discriminant learning in the acoustic domain. Our method aims to automatically determine from a large dimensional acoustic space a set of discriminant prosodic cues and/or patterns from a large dimensional acoustic space in a prominence detection task. In the present study, the authors propose to investigate two new assumptions:

- H_1: we support that the phenomenon of prominence results from a complex mechanism of interaction of different prosodic cues. This interaction takes place both in the combination of heterogeneous acoustic features (pitch, duration, vocal quality, etc.) and in the acoustic context-dependencies (combination of multiple reference floors according to multiple context-window sizes).

- H_2: we suggest that taking into account the phonetic-context, by neutralizing the variations related to the contextual pho-

netic properties (acoustic intrinsic properties and coarticulation effect), will improve the acoustics separation of prominent classes (P) and non-prominent (NP) syllables;

The article is organized as follows: first, we briefly describe the corpus used for the experiment; then, we describe the scheme of our proposed method; finally, we introduce our evaluation protocol and present the obtained results.

2. Speech Material

Unlike for English studies (Boston University Radio Speech Corpus [19], and Boston Directions Corpus [20]) there is currently no reference corpus available for prominence detection in French. Therefore a home-made corpus of a single speaker French read speech has been used for this preliminary study. The corpus has been manually annotated for prominence by 2 non-expert annotators with an agreement score of 78% of f-measure (respectively 80% of accuracy). This score is in concordance with the inter-annotator agreement usually observed in the literature (80-90% accuracy). For the detection task, only syllables for which agreement occurs were considered as being prominent; others have been declared non-prominent. This results in 1670 prominent syllables and 4635 non-prominent syllables.

3. Exploratory and Discriminant Learning Framework

3.1. Feature extraction

In [18], we have proposed a feature extraction framework for the prominence detection task. We outline here the main principles of the feature extraction step: we represent the acoustic space into 5 of his prosodic dimensions: pitch (f_0), durational (syllable duration and local speech rate), intensity (absolute and perceived), spectral (mel-frequency energy and loudness), and vocal quality (spectral slope, spectral emphasis, open quotient of the glottal source) features. On the basis of these prosodic dimensions, low-level statistical syllable-based features are extracted: minimum, maximum, mean, slope, glide, excursion, range, and standard deviation. These features are then used to give information on 2 different levels:

- intra-syllabic features: describe prosodic cues that occurs inside a given syllable;

- extra-syllabic features: describe the acoustic-context dependencies such as context-window and reference floor that account for acoustic-context.

In this experiment, two syllable-based units (syllable and nucleus of the syllable), several contextual-windows (none, {previous; next; surrounding } syllable, current breath group, current utterance) and two reference floors (minimum and mean value on the considered context-window) have been used.

3.2. Discriminant Analysis

The aim of our proposed method is to automatically estimate and accumulate the discriminant ability of the single prosodic cues of the considered acoustic space in order to improve performance in detecting prominence (H_1). To formalize this hypothesis, we set our approach within the framework of discriminant analysis. However, as the feature extraction step described above leads to a high dimensional feature space, this could causes several problems in a classification task. In particular: poor

classification performance due to the fact that some features are irrelevant for the task and therefore introduce noise in the learning procedure; over-fitting of the model to the learning set, especially in the case of dimension reduction algorithm such as LDA.

To prevent these bad properties, we propose the following learning scheme: 1) the initial feature space is reduced by feature selection into a subset of uncorrelated and discriminant features; 2) the resulting feature sub-space is transformed by means of linear discriminant analysis (LDA) in order to accumulate the discriminant ability of each single prosodic cue.

Let us introduce some notations and definitions. Let K be the total number of classes, N_k the number of total feature vectors accounting for the training data from class k and N the total number of feature vectors, μ_k and Σ_k respectively the mean vector and the covariance matrix of the class k, and μ the overall mean vector. The within and between class scatter is then defined as follows:

$$B = \sum_{k=1}^{K} N_k (\mu_{\mathbf{k}} - \mu)(\mu_{\mathbf{k}} - \mu)^T$$

$$W = \sum_{k=1}^{K} \sum_{n_k=1}^{N_k} (x_{n_k} - \mu_k)(x_{n_k} - \mu_k)^T$$

And the inertia ratio is defined as:

$$r = \frac{B}{W}$$

3.2.1. Feature Selection

The method used for the feature selection step is based on *Inertia Ratio Maximization using Feature Space Projection* (*IRMFSP*) [21, 18]. Let $\mathcal{F}$ be the acoustic feature space with respective feature dimensions $(f_i)_{1 \leq i \leq D}$ where D is the rank of the features space. At each iteration of the process, the best feature f_i is selected as the one maximizing r_i (inertia ratio along dimension i) and then the feature space is projected along the dimension f_i thus defining new dimensions f_i that verify:

$$\mathbf{f_j} = \mathbf{f_j} - (\mathbf{f_j} \cdot \frac{\mathbf{f_i}}{\|\mathbf{f_i}\|}) \frac{\mathbf{f_i}}{\|\mathbf{f_i}\|}, \forall \mathbf{f_j} \in \mathcal{F}$$

This projection step ensures that selected features are little corellated thus avoiding redundancy in the resulting features set. Iteration is processed until the ratio of the inertia ratio at step j to the inertia ratio of the first selected feature drop below a given threshold or when a desired number of features is reached. This leads to select a set of little-correlated features that individually maximizes the classes separability conditionally to each previous subset of selected features.

3.2.2. Feature Transform

The selected features estimated in the previous step are then used as input features of a linear discriminant analysis. The aim of this step is to combine the discriminant ability of the feature, by estimating the linear combinations of single features of the original space that maximize the class separation (following the same criterion mentioned above) This problems turns into determining the K-1 linear combinations $(\alpha)_{1 \leq k \leq K-1}$ that maximize the objective function:

$$J(\alpha) = \frac{\alpha^T B \alpha}{\alpha^T W \alpha}$$

Once the parameters α have been estimated, the original features are projected along these directions into a transformed space of rank K-1.

3.3. Context-Dependant Analysis

Even if context-dependent analysis has already been used by [12] in prominence detection, we would like to assume it as a full hypothesis in the acoustic modeling of prominence. In order to account for the effect of phonetic-context (intrinsic properties and coarticulation) on the observed prosodic parameters (H_2), we introduced syllable-based left-to-right contextual factors. The phonologic contextual factors used in this experiment are:

- type of {previous, current, next} syllable;

- number of phonemes in {previous, current, next} syllable;

- phonetic label of the current syllable's nucleus;

- phonetic class of the current syllable's nucleus;

- number of phonemes of {previous; current; next} syllable onset;

- number of phonemes of {previous; current; next} syllable coda;

- phonetic class of the current syllable onset & coda (i.e glide, occlusive, fricative, liquid, nasal);

4. Models evaluation

4.1. Compared Models

In this study, we compared 5 types of models of prominence detection. In order to remove any influence of the classifier type on the performance measurement, each experiment was performed using weka's J48 decision tree [22]. The compared models are:

- a baseline model (BL): decision tree learned on the whole acoustic space.

- a Context-Dependent model (H_1): decision tree learned on the whole acoustic space + phonetic-context features.

- an IRMFSP model: decision tree learned with the N first discriminant features selected with IRFMSP.

- an IRMFSP + LDA model (H_2): decision tree learned on the optimal linear combination of the N first discriminant features selected with IRMFSP.

- an IRMFSP + LDA + context-dependent ($H_1 + H_2$): decision tree learned on the optimal linear combination of the N first discriminant features selected with IRFMSP + phonetic-context feature

4.2. Evaluation scheme

The evaluation was conducted on the read speech corpus presented in section 2 within a 10-fold cross validation. Unlike other studies that use global accuracy as a measure of prominence detection performance, in this study we used the f-measure of the prominence class only. This choice is motivated by the following reasons: 1) prominence classification task is a detection task, i.e. a classification task in which only one of the class is of interest. This choice is all the more justified in the prominence detection task since the class of interest has much less observations than the other. 2) the f-measure provides a good compromise between recall and accuracy measures (insertion and deletion rates in the case of a two-class classification task).

5. Results and Discussion

In table 1, we summarize the results obtained with the 5 proposed models. Our proposed hypothesis (H_1: phonetic context-dependency and H_2: prosodic cues combination) leads to both significant improvement compared to the baseline method, with respective performance of 84.1 and 87.5% f-measure on the prominence class (this represents for comparison with standard performance measure respectively an accuracy of 86.5 and 90.4%).

Models	Mean Performance (% f-measure)
BL	81.5
H_1	84.1
IRMFSP	84.5
H_2	87.5
$H_1 + H_2$	87.4

Table 1: Performance of the compared models. In the case of the IRMFSP-based model, only the optimal model's performance is referred.
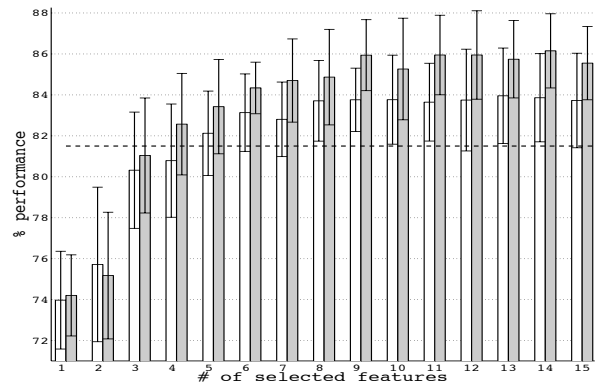

Figure 1: Performance of the IRMFSP (white bars) and the IRMFSP/LDA (gray bars) models as a function of the number of selected features in the feature selection step. Performances are presented as mean and standard deviation of the f-measure on the prominence class. The dotted line gives the mean performance of the baseline model. For convenience, performance is shown only for the first 15 selected features.

From the prosodic cue combination hypothesis (H_2), we have investigated several properties compared to the baseline as well as the IRMFSP models. First we have investigated the discriminant ability of the most discriminant prosodic feature (syllable duration) compared to the combination of the N most discriminant features (H_2). An anova analysis reveals that combination of prosodic features is significantly more discriminant for any number of selected features ($F-stats = 385$ for the most discriminant features vs. $511 \leq F-stats \leq 990$ for the combination of the N most discriminant features $N \geq 2$ on the test set) . A comparison of the IRMFSP and IRMFSP+LDA models by mean of prominence detection performance is presented in figure 1. Both models outperform the baseline model with a very small feature space dimension. The IRMFSP+LDA model outperforms the IRMFSP for almost all number of selected features ($N > 2$). This means that our proposed model succeed in finding discriminant prosodic patterns that are more discriminant than single cues estimated with the IRMFSP method. These results support the hypothesis that prosodic cues combine their contributions to achieve an acoustic saliency.

From the context-dependant hypothesis, it can be seen that the increase in performance observed with our H_1 model compared to the BL model and the combined $H_1 + H_2$ model compared to the H_2 model is relatively small or even null. This means that the proposed model somehow failed in catching discriminant phonetic distinctions in prominence structure. This suggests that the phonetic distinction does not affect overall prosodic features homogeneously; phonetic-context rather affects each particular prosodic cue depending on his nature (f_0, duration, intensity, spectral feature, voice quality). Thus, distinct context-dependent models should be learned for each prosodic cue or at least for each prosodic dimension. This remains to be investigated in a further study.

6. Conclusion and future works

In this paper we have proposed the investigation of two hypotheses for a syllable-based prominence detection model within a discriminant-based exploratory data analysis framework. These hypotheses were phonetic context-dependencies and prosodic cues combination. They both significantly outperform a baseline decision tree model in prominence detection. The improvement obtained by combining prosodic cues gives strong support to the hypothesis that prominence perception results from the combination of multiple prosodic cues contributions. However, our context-dependant model could be greatly improved by learning different context-dependant models according to the considered prosodic feature. Moreover, since the discriminant ability of a given prosodic cue is somehow related to his phonetic context, the context-dependant model should be estimated before estimating the discriminant ability of any prosodic cue; thus the context-dependency model and the feature selection should be merged into a single process. This preliminary study was achieved on read speech, further experiments will now be carried out on spontaneous speech corpora in order to propose a model robust to speaker, speaking style, expressivity and recording conditions. This will be carried out via the elaboration of a corpus of six hours of spoken French sampled into different speaking styles with manual prosodic annotations (*Rhapsodie* project). Furthermore, our proposed approach can be applied to any language since our model automatically estimates and combines relevant prosodic cues used for prominence achievement without any language-specific assumption. Thus it could be used in other languages to compare the language-specific prominence acoustic correlates and their relative weights in the achievement of an acoustic saliency.

7. Acknowledgments

8. References

[1] A. Erikson, E. Grabe, and H. Traunmuller, "Perception of syllable prominence by listeners with and without competence in the tested language," in *Speech Prosody*, Aix-en-Provence, France, 2002, pp. 275–278.

[2] P. Wagner, "Great expectations - introspective vs. perceptual prominence ratings and their acoustic correlates," in *Interspeech*, Lisbon, Portugal, 2005, pp. 2381–2384.

[3] J. Terken, "Fundamental frequency and perceived prominence," *J. Acoust. Soc. Am.*, vol. 89, pp. 1768–1776, 1991.

[4] P. Mertens, "Local prominence of acoustic and psychoacoustic functions and perceived stress in french," in *International Conference of Phonetic Science*, Aix-en-Provence, 1991, vol. 3, pp. 218–221.

[5] N. Campbell, "Loudness, spectral tilt, and perceived prominence in dialogues," in *ICPhS*, Stockholm, Sweden, 1995, pp. 676–679.

[6] A.M. Sluijter and V.J. Van Heuven, "Spectral balance as an acoustic correlate of linguistic stress," *Journal of the Acoustic Society of America*, vol. 100, no. 4, pp. 2471–2485, 1996.

[7] C. Gussenhoven, B.H. Repp, A. Rietveld, H.H. Rump, and J. Terken, "The perceptual prominence of fundamental frequency peaks," *Journal of the Acoustic Society of America*, vol. 102, no. 1, pp. 3009–3022, 1997.

[8] G. Fant and A. Kruckenberg, "Prominence correlates in swedish prosody," in *International Conference of Phonetic Science*, San Francisco, USA, 1999, vol. 3, pp. 1749–1752.

[9] A. Erikson, G.C. Thunberg, and H. Traunmüller, "Syllable prominence: A matter of vocal effort, phonetic distinctness and top-down processing," in *Eurospeech*, Aalborg, Denmark, 2001, pp. 399–402.

[10] M. Heldner, "On the reliability of overall intensity and spectral emphasis as acoustic correlates of focal accents in swedish," *Journal of Phonetics*, vol. 31, no. 1, pp. 39–62, 2003.

[11] A. Rosenberg and J. Hirschberg, "Detecting pitch accent using pitch-corrected energy-based predictors," in *Interspeech*, Antwerp, Belgium, 2007, pp. 2777–2780.

[12] K.L. Jenkin and M.S. Scordilis, "Development and comparison of three syllable stress classifiers," in *ICSLP*, 1996, vol. 2, pp. 733–736.

[13] Tamburini F., "Automatic detection of prosodic prominence in continuous speech," in *International Conference on Language Resources and Evaluation (LREC'2002)*, Las Palmas, Canary Islands, Spain, 2002, pp. 301–306.

[14] X. Sun, "Pitch accent prediction using ensemble machine learning," in *ICSLP*, Denver, USA, 2002, pp. 953–956.

[15] G. Kochanski, E. Grabe, J. Coleman, and B. Rosner, "Loudness predicts prominence: fundamental frequency lends little," *J. Acoust. Soc. Am.*, vol. 118, pp. 1038–1054, 2005.

[16] S. Ananthakrishnan and S. Narayanan, "An automatic prosody recognizer using a coupled multi-stream acoustic model and a syntactic-prosodic language model," in *International Conference on Acoustics, Speech, and Signal Processing*, 2005.

[17] M. Avanzi, J.-P. Goldman, A. Lacheret-Dujour, A.-C. Simon, and A Auchlin, "Méthodologie et algorithmes pour la détection automatique des syllabes proéminentes dans les corpus de français parlé," *Cahiers of French Language Studies*, vol. 13, no. 2, 2007.

[18] N. Obin, X. Rodet, and A. Lacheret-Dujour, "Prominence model: a probabilistic framework," in *International Conference on Acoustics, Speech, and Signal Processing*, Las Vegas, NV, USA, 2008, pp. 3993–3996.

[19] M. Ostendorf, P. Price, and S. Shattuck-Hufnagel, "The boston university radio news corpus," 1995.

[20] C. Nakatani, J. Hirschberg, and B. Grosz, "Discourse structure in spoken language: studies on speech corpora," in *Working Notes of AAAI-95 Spring Symposiom on Empirical Methods in Discourse Interpretation*, 1995.

[21] G Peeters, "Automatic classification of large musical instrument databases using hierachical classifiers with inertia ratio maximization," in *AES 115th Convention*, 2003.

[22] I. Witten, E. Frank, L. Trigg, M. Hall, G. Holmes, and S. Cunningham, "Weka: Practical machine learning tools and techniques with java implementation," in *ICONIP/ANZII,ANNES*, 1999, pp. 192–196.