FRENCH PROMINENCE: A PROBABILISTIC FRAMEWOK

Nicolas Obin, Xavier Rodet

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

Anne Lacheret-Dujour

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

ABSTRACT

Identification of prosodic phenomena is of first importance in prosodic analysis and modeling. In this paper, we introduce a new method for automatic prosodic phenomena labelling. The authors set their approach of prosodic phenomena in the framework of prominence. The proposed method for automatic prominence labelling is based on well-known machine learning techniques in a three step procedure: i) a feature extraction step in which we propose a framework for systematic and multi-level speech acoustic feature extraction, ii) a feature selection step for identifying the more relevant prominence acoustic correlates, and iii) a modelling step in which a gaussian mixture model is used for predicting prominence. This model shows robust performance on read speech (84%).

Index Terms— Prosody, prominence, acoustic correlates, feature selection, classification, gaussian mixture model

1. INTRODUCTION

The identification of prosodic phenomena is an essential task in the analysis of prosody as well as for its modeling in the context of text-to-speech systems. Understanding acoustic correlates of these phenomena in order to automatically detect them from speech is of great help for prosodic models. Recent automatic prosodic annotation research has focused on prominence instead of accent [1, 2, 3]. In this paper, we present a prominence identification method based on a statistical model that enables the automatic emergence of acoustic correlates of prominence and then their automatic classification. This paper is organized as follows: firstly, we define the notion of prominence and how this is favorable to the concept of accent. In the second section, we clarify the protocol for the manual annotation of a reference corpus. In the third section, we explain the probabilistic framework based on well-known pattern matching methods: feature extraction, feature selection, and bootstrap learning method for prominence modeling with a gaussian mixture model.

2. WHAT IS PROMINENCE?

The transcription of prosodic phenomena is usually carried out using the notion of accentuation. Several systems for the transcription of prosody (ToBI [4] and RFC [5] for English annotation; INTSINT [6, 7] and [8] for French annotation) are based on this notion. This strategy takes *a priori* theoretical knowledge for granted and supposes an already-known phonological representation as well as its acoustic correlates and the associated prototypes. This definition of prosodic phenomena has several drawbacks: firstly, it supposes that the phonological system is already known, meaning their acoustic correlates, their type and their associated function are known. However, such phonological representations are not unanimous and the resulting annotations show large interindividual variations that contradict the strength of these models [9]. Recent studies have favored the notion of prominence over that of accent. By prominence, we refer to the definition stated in [10]: "prominence is the property by which linguistic units are perceived as standing out from their environment". In this paper, we will use the methodology defined in [3]: prominence is a perceptive phenomena that does not refer to a phonological system and of which one does not presuppose the acoustic correlates, nor the arrangement of the spoken chain. The considered prominent unit here is the syllable.

	NP	P	Total
NP	3385	543	3928
P	707	1670	2377
Total	4092	2213	

Table 1. Confusion matrix for P/NP decision task

3. PROMINENCE ANNOTATION

Prominence being a perceptive phenomenon, the first step of its modeling is the creation of a reference corpus based on a manual annotation. We have defined the following annotation protocol: two non-specialist individuals were simultaneously annotating a single speaker corpus of 466 read sentences containing 6305 syllables in sentences ranging from 2 to 66 syllables, with an average and standard deviation of 13.5 and 9.5 syllables. The annotation task was defined as follows: in each sentence, subjects were asked to note the group of syllables "P" for prominent or "NP" for non-prominent. Subjects could listen to each sentence as many times as they wished and using different temporal scales before making their decision. We present in table 1 the confusion matrix between the two annotators.

We define the agreement measure as being the mean of prominent and non-prominent f-measure. This measure i) gives compromise on recall and accuracy measures and ii) neutralizes the datas proportion effect (roughly 64% non-prominent and 36% prominent syllables). The result demonstrates agreement in discriminating prominent / non-prominent syllables (78.6% mean f-measure) and validates the concept of prominence as a robust perceptive correlate for a prosodic phenomena annotation task. Only syllables which show agreement during annotation were set as prominent for the rest of the paper.

4. ACOUSTIC CORRELATES OF PROMINENCE

Recent research shows that prosodic phenomena result from the interactions of acoustic cues that are more complex than pitch and duration. Local speech rate [11], loudness [2], and subbands energy [12] should be taken into consideration for such phenomena analysis. These studies indicate that prosodic acoustic correlates should not be restricted *a priori*. At the same time, prominence detection methods focusing on the search of new acoustic correlates are still based on arbitrary feature subsets: f_0 and duration [3]; f_0, duration and energy [1]; f_0, phone duration, loudness, aperiodicity and spectral slope [2], and f_0, duration and subbands energy [12]. This section introduces a systematic framework for the extraction of features with the goal of determining the acoustic correlates of prominence without *a priori* knowledge.

4.1. Methodology for acoustic features extraction

We propose to define a systematic framework for the description of speech acoustic properties as follows: i) statement of primitive acoustic features, ii) measurement of characteristic values for each feature over a given syllable, and iii) comparison of the considered syllable features according to different surrounding syllables temporal horizon.

4.1.1. Primitive acoustic features

The first step lies in the choice of the speech primitive acoustic features computed from the signal: pitch (fundamental frequency or f_0), duration features (syllable duration, nucleus duration, local speech rate [11] , intensity (energy and loudness), and spectral features (voiced/unvoiced cutoff frequency, spectral centroid, spectral slope, specific loudness).

4.1.2. Definition of characteristic values

The second step consists of defining the measurements that enable the description of these features over a given temporal segment - here the syllable. We defined three distinct groups of measurement:

- global characteristics: maximum value, minimum value, mean value, value sumation over unit,
- dynamic characteristics which give rough information on feature movement in the considered temporal segment: range and start to end value difference,
- shape features: first and second polynomial approximation, legendre polynomial approximation, 3rd order splines, *hu moment* and *zernike moment*). The last two features are derived from image shape analysis and have been added for their property of scale invariance, which appears to be convenient for prosodic shape clustering.

4.1.3. Analysis of multi-level features

As we have said previously, prominence is not only defined by intrinsic properties. It is essentially characterized as a salience in relation to syllables that surround it. Today, the temporal horizon of prominence processing has not been defined in publications. We suggest to heuristically define different temporal horizons for the comparison of acoustic data relevant for prominence detection. We have organized these temporal horizons into a hierarchy from the smallest to the largest. The characteristic values calculated over a given syllable segment are compared to those of: i) adjacent syllables (previous, following, and mean of both previous and following), ii) accentual group including the current syllable (excluding itself), iii) prosodic group including the current syllable (excluding itself), and iv) a sentence including the current syllable (excluding itself). We define the accentual group as being the segment between two consecutive prominences and the prosodic group as a set of accentual group followed by a silence. The accentual group was only considerd during the feature selection step where prominence annotation were available; it was set equal to the prosodic group when such information were not available. Such multi-level comparisons are illustrated in Figure 1 for fundamental frequency.

4.2. Acoustic correlates of prominence with a feature selection algorithm

Our feature extraction protocol results in the extraction of 1490 features. These features are obviously not all of equal importance according to prominence. We therefore propose to find the subset of features that best explain prominence phenomena. Our strategy for identifying these features from the complete feature set (as defined in the previous section) is based on a feature selection method. The goal of feature selection methods is basically to derive an optimal subset of features from an initial set following a given criterion.

The proposed method is based on *Inertia Ratio Maximization using Feature Space Projection* [13]. This method is based on iteratively finding the feature that maximizes the *Inertia Ratio* and then projecting the data orthogonally to this feature. Let K be the total number of classes - here 2, prominent and non-prominent -, N_k the number of total feature vectors accounting for the training data from class k and N the total number of feature vectors. Let X_{i,n_k} be the n_k-th feature vector along dimension i from the class k, $m_{i,k}$ and m_i respectively the mean of the vectors of the class k $(X_{i,n_k})_{1<i<N_k}$ and the mean of all training vectors $(X_{i,n_k})_{1<i<N_k,1<k<K}$.

The Inertia Ratio is defined as the ratio of the Between-class-inertia B_i to the average radius of the scatter of all classes R_i:

$$r_i = \frac{B_i}{R_i} = \frac{\sum_{k=1}^{K} \frac{N_k}{N} \|m_{i,k} - m_i\|}{\sum_{k=1}^{K} (\frac{1}{N_k} \sum_{n_k=1}^{N_k} \|x_{i,n_k} - m_{i,k}\|)} \quad (1)$$

The method is iterative: at each step, the selected feature i_{opt} is the one which maximizes the Inertia Ratio. Then features are orthogonally projected along the i_{opt} feature. This projection step ensure non-redundancy in the selected features subset. Before computing feature selection, feature vectors were first normalized according to their standard deviation over the class k. This treatment normalizes distance measures during the feature selection procedure.

In Table 2, we summarize the 10 most relevant prominence acoustic features in order of relevance. The main relevant features of prominence are: duration features (syllable duration, local speech rate and nucleus duration), pitch feature (f_0), and spectral features (specific loudness). This result does not validate results for loudness predominance [2] in case of french prominence. Secondly, it indicates that prominence perception results from a complex interaction of features: this phenomen involves absolute features as well as relative features over different temporal horizons (previous syllable, next syllable, accentual group), as well as shape features. It can be somewhat surprising that shape features do not appear there, when f_0 shape is expected to be a relevant feature for prominence identification. This could be explained for two reasons: i) f_0 shape is not relevant for prominence identification, ii) the shape representations chosen here are not able to catch what distinguish prominent from prominent f_0 shape. This remains to be investigated.

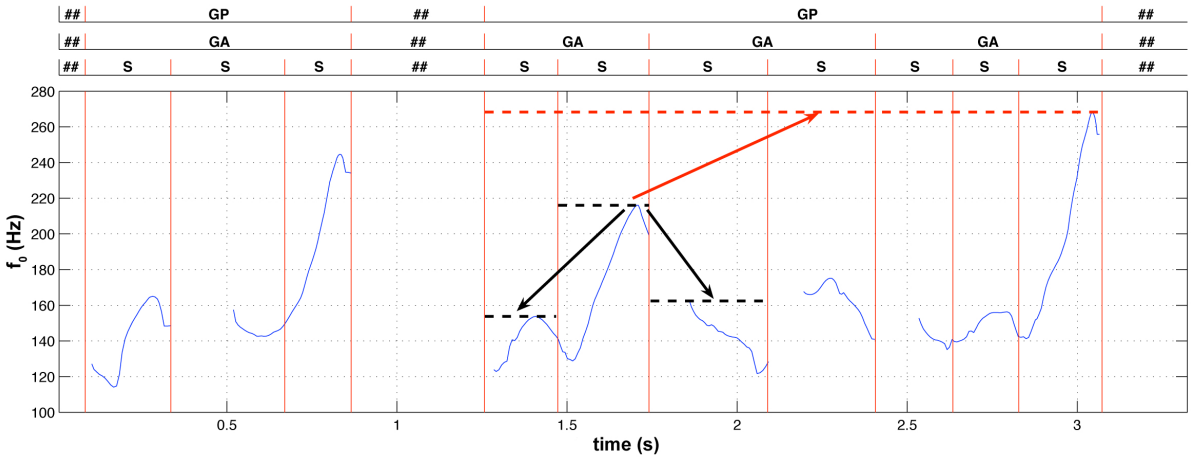

Fig. 1. Comparison of fundamental frequency local maxima over several temporal horizons: the local maxima over syllable is compared with local maxima over adjacent syllables and parent prosodic group.
(S: Syllable segment, GA: Accentual Group and GP: Prosodic Group.)

Feature name	r_i
duration	37.46
f_0 mean / around f_0 mean	5.81
nucleus duration	4.65
spec loud 1 mean / next spec loud 1 mean	1.68
spec loud 18 mean / prev spec loud 18 mean	1.36
lsr min / ag lsr min	1.32
lsr curve	0.90
lsr mean / around syl lsr mean	0.59
energy mean	0.51
lsr slope	0.40

Table 2. Selectionned features according to IRMFSP with respective inertia ratio. For clarity, spec loud = specific loudness with respective band, lsr = local speech rate. When not mentionned the temporal horizon is the syllable; ag is the accentual group

5. PROMINENCE MODEL

Once the most relevant acoustic correlates on the reference corpus have been determined, we need to model the prominent and non-prominent classes for class prediction. We chose the well-known *Gaussian mixture model* (GMM) for data modeling.

5.1. Gaussian mixture model

For each prominent and non-prominent class, the distribution of the P-dimensional feature vectors is modeled by a Gaussian mixture density. Then for a given feature vector x, the mixture density for class k is defined as:

$$P(x|k) = \sum_{i=1}^{M} \omega_k^i b_k^i(x) \qquad (2)$$

where the weighting factors ω_k^i are positive scalars satisfying $\sum_{i=1}^{M} \omega_k^i = 1$. The density is then a weighted linear combination of M gaussian densities b_k^i with mean vector μ_k^i and covariance matrix Σ_i^k. The model parameters $\theta_k = \{\mu_k^i; \Sigma_k^i; \omega_k^i\}_{i=1,\ldots,M}$ are estimated with the *Expectation-Maximization* (EM) algorithm [14]. Classification is then made using the *Maximum a posteriori Probability* (MAP) decision rule. Models were trained with the first 100 features issued from the preceding feature selection step (section 4).

5.2. Learning procedure

Our proposed learning method is a two-step method: in a first supervised step, model parameters $\theta_{k,0}$ are estimated on a reference corpus for the prominent and non-prominent classes. These parameters are used as initialization in an iterative unsupervised prediction/learning procedure. Given an iteration i of the method, a class label sequence is first estimated according to the MAP decision with the previous models $\theta_{i-1} = \{\theta_{j,i-1}\}_{j=1,\ldots,K}$. Then, model parameters are reestimated for each class k according to the predicted class label sequence with initialisation model $\theta_{k,i-1}$ and prior probability equal to posterior probability of the $\theta_{k,i-1}$ model. This reestimation of the model parameters gives the model θ_i. Iteration is computed until model convergence.

For this procedure we have built three corpora for model initialization, learning, and validation steps. Firstly, the reference corpus has been equally split into an initialization corpus and a validation corpus.Secondly, a non-annotated corpus was used for the unsupervised model learning step. This last corpus contains 69688 syllables distributed into 3615 sentences from 2 to 74 syllables with a mean and standard deviation respectively of 19 and 9 syllables.

We define the performance measure as the mean of prominence and non-prominence fmeasure - this for the same reasons that stated in Section 3.

Initialization and validation corpora have both been used for performance measures: the perfomance on initialization corpus indicates the learning ability of our model, whereas performance on the validation corpus indicates generalization ability. The performance measure is computed at each step of the learning procedure. Different mixture componants have been tested on the same procedure from 2 to 16 components; as well as different learning corpus sizes equally spaced from 20% to 100% of the whole corpus. Finally, initialization and validation corpora were inverted in a cross-validation procedure.

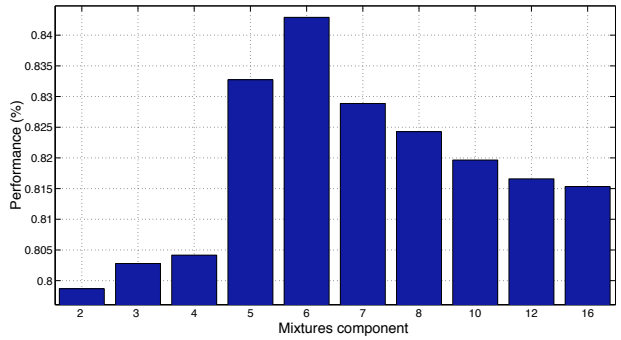

Fig. 2. Mean performance in prominence detection as a function of mixtures component. The mean was computed on the performance according to data ratio in learning corpus.

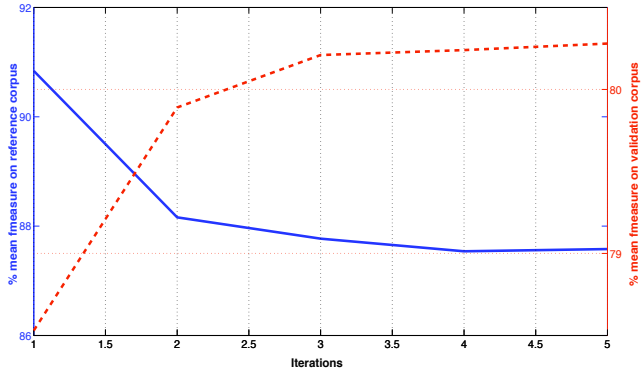

Fig. 3. Evolution of mean performance during unsupervised learning. In plain line, mean recall on initialization corpus and in broken line, mean recall on validation corpus

5.3. Results and discussion

Figure 2 summarizes our model mean performance results. Our model has an overall mean performance of 83% on the initialization corpus and 80% on the validation corpus. Maximum performace is of 89% on the validation corpus, which is encouraging result. The optimal model was found to be a 6 components mixture with 84% mean generalization performance and 89% maximum generalization performance. Then the performance decreases with model order as the model starts to overfit data and loses generalization ability. Figure 3 presents an example of the evolution of performance as a function of the iteration step during unsupervised learning. The model improves generalization performance since learning performance decreases. This means that i) the model is learning prominence structure on the unknown dataset and ii) the model is learning general prominence characteristics instead of corpus dependant ones. The cross-validation procedure gives comparable performances with 85% on the initialization corpus and 82% on the validation corpus. This means that the learning procedure is robust since model performance does not depend on the initialization dataset.

6. CONCLUSION AND FUTURE WORKS

We have shown with a feature selection algorithm that prominence phenoma results from a complex interaction of acoustic correlates. Our proposed model for automatic prominence prediction shows good and robust performances. In the feature selection step, it could be expected that prominence relies on multiple and specific acoustic cues which define the so-called prominence type. This stated, the feature selection should account for such type characteristics by preprocessing the prominence class with clustering methods or directly with unsupervised feature selection methods. In the modelling step, the choice of both feature number and classifier for detection were arbitrary set. In future research, we are interested in i) investigating the effect of both feature number and classifier type on the detection performance in order to estimate an optimal model for prominence detection, ii) modelling acoustic prominence type with clustering methods, iii) defining a prominence strengh measure that would be used for prosody modeling and prediction from text structure.

7. REFERENCES

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

[2] 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.

[3] 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.

[4] M. Beckman, J. Hirschberg, and S. Shattuck-Hufnagel, "The original tobi system and the evolution of the tobi framework," in *Prosodic models and transcription: Towards prosodic typology*, pp. 9–54. Oxford University Press, Oxford, 2004.

[5] P. Taylor, "The rise/fall/connection model of intonation," *Speech Communication*, vol. 15, no. 1-2, pp. 169–186, 1994.

[6] D. Hirst, N. Ide, and J. Veronis, "Coding fundamental frequency patterns for multi-lingual synthesis with intsint in the multext project," in *Proceedings of the 2nd ESCA/IEEE Workshop on Speech Synthesis*, New Paltz, NY, 1994, pp. 77–81.

[7] B. Post, E. Delais-Roussarie, and A-C. Simon, "Ivts, un système de transcription pour la variation prosodique," *Bulletin PFC*, vol. 6, pp. 51–68, 2006.

[8] S-A. Jun and C. Fougeron, "A phonological model for french intonation," in *Intonation: Analysis, Modeling and Technology*, A.Botinis, Ed., pp. 209–242. Kluwer, Dordrecht, 2000.

[9] C. Wightman, "Tobi or not tobi?," in *Proceedings of the First International Conference on Speech Prosody (SP'2002)*, Aix-en-Provence, France, 2002, pp. 25–29.

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

[11] H. Pfitzinger, "Local speech rate as a combination of syllable and phone rate," in *Proc. of International Conference on Speech Language Processing (ICSLP'1998)*, Sydney, Australia, 1998, vol. 3, pp. 1087–1090.

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

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

[14] Dempster A.P., Laird N.M., and Rubin D.B, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society*, vol. B, no. 39, pp. 1–38, 1977.