French Prominence, a probabilistic framework

Anne Lacheret, Nicolas Obin, Xavier Rodet

FRENCH PROMINENCE: A PROBABILISTIC FRAMEWOK

N. Obin[*], *X. Rodet*

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

A. 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. We introduce in this paper a new method for automatic prominence labelling. The proposed method is based on well known machine learning technics in a three step procedure : feature extraction, feature selection for finding the more relevant prominence acoustic correlates and gaussian mixture model for predicting prominence.

Index Terms— Prosody, prominence, acoustic correlates, feature selection, 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. Understand acoustic correlates of these phenomena in order to automatically detect them from speech is of great help for prosodic models. Recent automatic prosodic annotation researches have 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 prominence acoustic correlates 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 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 the function of the prominences, their acoustic correlates, and their type 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.

3. PROMINENCE ANNOTATION

3.1. Methology

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 annotated simultaneously a single speaker corpus of 466 sentences containing 6185 syllables in sentences ranging from 2 to 66 syllables, with an average and standard deviation of 13 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. The subject could listen to each sentence as many times as he/she wished and using different temporal scales before making their decision. We present in table 1 the confusion matrix between the two annotators.

[*]The first author would like to thank B. Yegnanarayana for his very constructive suggestions.

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

Table 1. Confusion matrix for P/NP decision task

These results demonstrate agreement in discriminating prominent / non-prominent syllables (78.2% mean recall) and validates the concept of prominence as a robust perceptive correlate for a prosodic phenomena annotation task.

4. ACOUSTIC CORRELATES OF PROMINENCE

Recent research shows that prosodic phenomena result from acoustic cues interaction that are more complex than pitch and duration. Local speech rate [11], which indicates suprasegmental duration coherence, and loudness [2], should be taken into consideration for such phenomena analysis. Even if these acoustic features remain little studied in the prosodic fields, these studies indicate that prosodic acoustic correlates should not be set *a priori*. At the same time methods of automatic prominence detection still suppose a priori knowledge of the more relevant prominence acoustic features: f_0 and duration [3]; f_0, duration and energy [1]; f_0, phone duration, loudness, aperiodicity and spectral slope [2]. This section introduces the methodology used for the generation of features with the goal of determining the acoustic correlates of prominence without a priori knowledge.

4.1. Methology for acoustic features extraction

We propose to define a systematic framework for the characterization of different properties of speech as follows: statement of primitive acoustic features, measurement of characteristic values for each feature over syllable segments, and supra-segmental comparison of characteristics value calculated over each syllable segments. The first step is the consideration of the speech primitive acoustic features deduced from signal: pitch (fundamental frequency or f_0), duration features (syllable duration, nucleus duration, local speech rate [11] and the duration difference to the phonological syllable duration learned with decision tree on syllable internal structure), intensity (energy and loudness), and spectral features (voiced/unvoiced frequency, spectral centroid, spectral slope, specific loudness). The second step consists of the definition of the measurements that make it possible to characterize these features on a given temporal segment - here the syllable. We defined three groups of qualitative measurement: global characteristics (maximum value, minimum value, mean value, value sumation over unit), dynamic characteristics (range and start to end value difference) that give rough information on feature movement in the considered temporal segment, and 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.

As we have seen, prominence is not only defined by intrinsic properties. It is essentially characterized as a salience in relation to those that surround it. Today, the temporal horizon of prominence processing has not been defined in publications. We suggest heuristically defining different temporal horizons for the comparison of acoustic data relevant for prominence decision. 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: adjacent syllables (previous and following), accentual group including the current syllable (but with the exclusion of it), prosodic group including the current syllable (but with the exclusion of it), and a sentence including the current syllable (but with the exclusion of it). Such comparisons are illustrated on Figure 1 for fundamental frequency. The local maxima over syllable is compared with local maxima over adjacent syllables and parent prosodic group.

4.2. Acoustic correlates of prominence with feature selection algotihm

Our feature generation protocol results in the extraction of 1490 features. These features are obviously not of equal importance according to prominence. We propose then to find the subset of features that best explain prominence phenomena. Our proposed strategy for identifying those from the complete feature set that we defined in previous sections is based on a feature selection method. The goal of feature selection methods is 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 [12]. This method is based on iteratively finding the feature that maximizes the Fisher Discriminant and then projecting the data orthogonaly 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 Fisher discriminant is defined as the ratio of the Between-class-inertia B_i to the average radius of the scatter of all classes R_i :

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

The method is iterative : at each step, the selected fea-

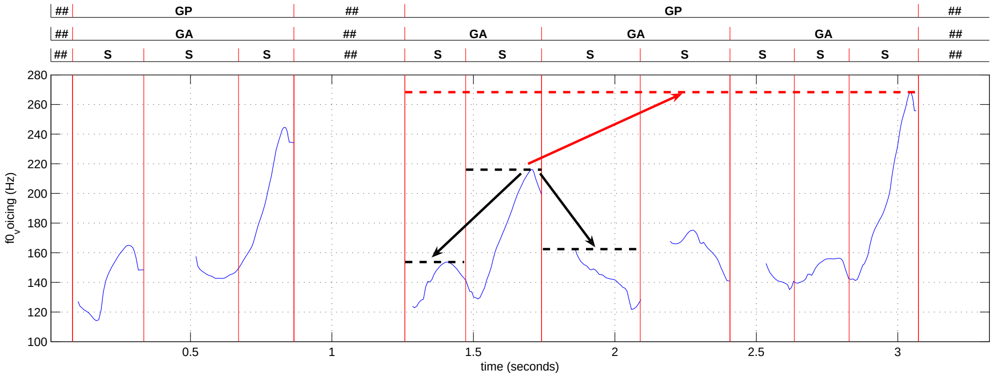

Fig. 1. Comparison of fundamental frequency local maxima over several temporal horizons. (S: Syllable segment, GA: Accentual Group and GP: Prosodic Group.)

ture i_{opt} is the one which maximizes the Fisher Discriminant. Then features are orthogonaly projected along the i_{opt} feature. This projection step ensure non-redundancy in the selected features subset. Features vector are normalizated according to their standard deviation over the class k. This treatment normalizes distance measures during the feature selection procedure.

According to this method the first ten more relevant prominence acoustic features appears to be (in order of relevance): syllable duration, local speech rate minimum, ratio of current syllable f_0 mean to previous syllable f_0 mean, syllable nucleus duration, ratio of current syllable 15th band specific loudness maximum to previous syllable 15th band specific loudness mean, ratio of current syllable f_0 mean to next syllable f_0 mean, 3rd order local speech rate spline model over accentual group, ratio of current syllable 2nd band specific loudness minimum to previous syllable 2nd band specific loudness mean, nucleus f_0 3rd order spline model and ratio of current syllable local speech rate minimum to previous syllable local speech rate mean.

This result first shows the prominence main relevant features : duration features (syllable duration, local speech rate and nucleus duration), pitch feature (f_0), and spectral features (specific loudness). This result does not validate [2] results for loudness predominance in case of french prominence . Secondly it indicates the complex features interaction in prominence perception: absolute features as well as relative features with different temporal horizons (previous syllable, next syllable, accentual group), and shape features.

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:

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

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 $\theta_k = \{\mu_k^i; \Sigma_k^i; \omega_k^i\}_{i=1,\dots,M}$ are estimated with the Expectation-Maximization (EM) algorithm. Classification is then made using the Maximum a posteriori Probability (MAP) decision rule. Models were trained with the first 100 features issued from the precedent 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 two classes prominent and non-prominent. Then 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 MAP decision with the previous models $\theta_{i-1} = \{\theta_{j,i-1}\}_{j=1,\dots,K}$. Then model parameters are reestimated for each class k according to the predicted class labels 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 build three corpus for model initialization, learning and validation steps. Firstly, the refer-

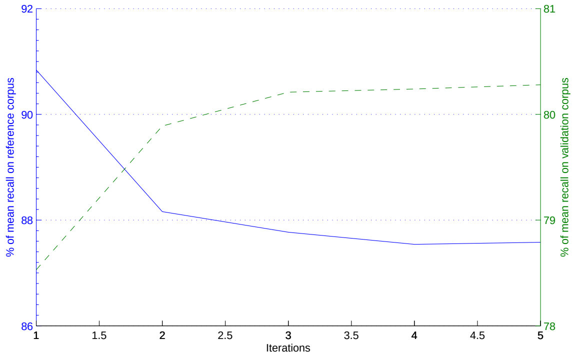

Fig. 2. Evolution of mean performance during unsupervised learning. In plain line, mean recall on initialization corpus and in broken line, mean recall on validation corpus

ence corpus has been equally split into an initialization corpus and a validation corpus. Secondly a non-annotated corpus was used for the unesupervised 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 defined the performance measure as the class recall mean of the confusion matrix between predicted and annotated classes.

Initialization and validation corpus have both been used for performance measures: the perfomance on initialization corpus indicates the learning ability of our model; when 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 equaly spaced from 20 % to 100 % of the whole corpus. Finally, intialization and validation corpus were inverted into a cross-validation procedure.

5.3. Results and discussion

Our model has an overall mean performance of 83% on initialization corpus and 80% on validation corpus. Maximum performace is of 98% on initialization corpus and 89% on validation corpus, which are encouraging results. Optimal model was found to be a 5 componants mixture with 85% mean generalization performance. Then the performance is decreasing with model order since the model starts to overfit data and loses generalization ability. Figure 2 presents the evolution of performance as a function of the iteration step during unsupervised learning. The model improves generalization performance since learning performance decreases. This firstly means that model is learning prominence structure on the unknown dataset and secondly that the model is learning general prominence characteristics instead of corpus dependant ones. The cross-validation procedure gives comparable performances with 85% on initialization corpus and 82% on validation corpus. This means that the learning procedure is robust since model performance does not depend on the intialization dataset.

6. CONCLUSION AND FUTURE WORKS

We have shown with a feature selection algorithm that prominence phenoma results of acoustic correlates complex interaction. Our proposed model for automatic prominence prediction shows good and robust performances. As a future work, feature selection and prominence modeling should be merged into a single procedure with a bayesian network. Unsupervised clustering methods should also be used for acoustic prominence type modeling. Then a prominence strength measure should be defined 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] G Peeters, “Automatic classification of large musical instrument databases using hierachical classifiers with inertia ratio maximization,” in *AES 115th Convention*, 2003.