Sentence Modality Recognition in French based on Prosody

Pavel Král[1], Jana Klečková[1], Christophe Cerisara[2]

[1]Dept. Informatics & Computer Science
University of West Bohemia
Plzeň, Czech Republic
[2]LORIA UMR 7503
BP 239 - 54506 Vandoeuvre
FRANCE
kral,kleckova@kiv.czu.cz,cerisara@loria.fr

Abstract—This paper deals with automatic sentence modality recognition in French. In this work, only prosodic features are considered. The sentences are recognized according to the three following modalities: declarative, interrogative and exclamatory sentences. This information will be used to animate a talking head for deaf and hearing-impaired children. We first statistically study a real radio corpus in order to assess the feasibility of the automatic modeling of sentence types. Then, we test two sets of prosodic features as well as two different classifiers and their combination. We further focus our attention on questions recognition, as this modality is certainly the most important one for the target application.

Keywords—prosody, fundamental frequency (F0), energy, automatic sentences modality recognition (ASMR), modal corpus.

1. INTRODUCTION

This work aims at developing applications to help deaf and hearing-impaired children to better understand and be integrated in classrooms with normal-hearing children. Here, we investigate the possibility to recognize sentence modality (questions or exclamations) in French from prosodic features.

The software that the children may use in the future is based on the following principle: a microphone captures the speech signal of the teacher, which is then passed to a phonetic speech recognizer. The sequence of phones recognized by the system is then translated into "Langage Parlé Completé" (LPC) [1], which is a visual representation of the phonetic content of the sentence. This representation, well-known by part of the deaf community, is based on lip movements enriched by hands and fingers positions. In the laptop used by a child, a 3D talking head [2] reproduces these lip and hand movements.

The objective of this work is to study the possible use of prosodic features to automatically recognize three classes of sentence types: questions (Q), exclamations (E) and declarative sentences (D). Such an information may then be used to enrich the LPC transcription that appears on the laptop screen, for example by displaying a question mark near the talking head and by raising the brows of the 3D head when a question is asked.

Obviously, different types of information (prosody, syntax, semantic, ...) shall be used to recognize sentence modality. In this work, we are only concerned by two prosodic features: fundamental frequency (F0) curve and energy. The next versions of the system will integrate other knowledge sources. It is also important to note that, in the context of the above described application, questions are the most important types of sentences to detect. Therefore, in this paper, a particular attention is given to questions.

2. SHORT REVIEW OF MODALITY RECOGNITION APPROACHES

The basic rules concerning French sentences prosody can be summarized as [3]:

- Declarative sentence: small decrease of melody,

- Imperative sentence: important decrease of melody,
- Interrogative sentence: increase of melody,
- Grammar interrogative sentence: neutral intonation.

French prosody is studied in a number of fields, for example in emotion recognition [4], but very few papers use it for automatic modality recognition. On the other hand, sentence type recognition is much more studied in other languages and particularly in English. In the published works, the following features are used:

- F0 contour in [5] for German,
- F0 and energy in [6] for German and English,
- F0 and energy in [7] for Czech,
- F0 and duration of the ending suffix in [8] for standard Korean.

Another work [9] investigates many other prosodic attributes that are mostly derived from F0, energy and duration, such as the max, min, mean and standard deviation of F0, the energy mean and standard deviation and the number of frames in utterance and number of frames of F0. The features are computed on the whole sentence and also on the last 200ms of each sentence. The authors conclude that the end of sentences is the most important part for modality recognition.

In the litterature, the following classification methods have been tested and compared for sentence type recognition: Neural Networks (NN) [5, 7, 10], Hidden Markov Models (HMMs) [9] and Classification and Regression Trees (CART) [9, 10]. The error rate is comparable between such classifiers.

3. MODAL CORPUS BUILDING

The final system should be trained on a corpus recorded in real classrooms, but such a corpus is not available for now. Therefore, we looked for an existing French corpus annotated with sentence modalities that would fit our needs, and the less problematic we found is the ESTER corpus [11], used in the French broadcast news evaluation. As this corpus has not been designed *a priori* to do sentence modality recognition, we decided to relabel it for this task.

Manually labeling sentence modality is very subjective, and different labels are often given by different persons for the same sentence. Therefore, we can identify actually 7 possible modalities for each sentence, as shown in figure 1, instead of the three original ones.

The objective of the modality recognizer shall first be clearly defined: it may be for example to classify as X all the sentences that *can* be annotated as X. In our case, we have rather chosen to recognize only the "non-overlapping" subsets Q̌ Ě and Ď, and we have thus built three distinct models, one for Q̌, one for Ě and one for Ď.

We used the three punctuation marks "? . !" to extract from the raw ESTER corpus a set of sentences that belongs to each category. This first modal corpus (hereafter called *original* corpus) contains 927 sentences (324 declarations, 351 exclamations and 252 questions) for training and 429 sentences (150 declarations, 153 exclamations and 126 questions) for testing.

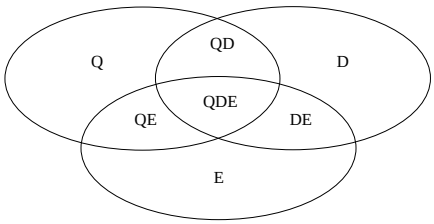

Fig. 1. Seven modal classes according to sentence types labeling: X labels: the sentences considered as X by all labelers; XY labels: the sentences considered as X by some labelers and as Y by some others; XYZ labels: the sentences that belong to the three main classes, depending on the labeler.

But such punctuation marks are only indicative and do not represent accurate sentences modality. We can thus assume that the "?" punctuation actually represents the "broad" overlapping Q class represented in figure 1. To obtain the "Q̌-only" class representatives, we have first extracted the "?" sentences, and then re-labeled these sentences as Q̌ or non-Q̌, where Q̌ only contains the sentences that are *surely* questions. Of course, this manual labeling still contains errors, but it is clearly better than the original one. This manual "filtering" has been realized for all three classes on the whole test corpus, but only on half of the training corpus, because of its size. The resulting corpus is hereafter called *filtered* corpus.

4. EXPERIMENTS

We first study the caracteristics of the F0 curve at the end of the sentence. The basic prosodic rules in the French language assume that F0 increases for Q, decreases for E and is quasi-stationary for D [3]. Even though such rules are obviously only crude approximations of practical French prosody, we try through the following experiments to assess the importance of such an attribute for sentences modality recognition.

4.1. Statistical study of the modal corpora

We first perform a statistical study of the original corpus. The objective of this study is to compare the distribution of the final F0 slope in each category (Q, E, D). We thus compute the slope of F0 during the last 0.7s of speech for each sentence, train and test together. This is done by first

estimating 4 values of F0 using the autocorrelation function as described in [12], and then applying a linear regression on these 4 values. The distributions are shown in the left column of figure 2. The top, median and bottom histograms respectively represent the distributions of exclamations, declarative sentences and questions. We can observe that the final F0 slope does not clearly discriminate between the three modalities.

Next, the same analysis is performed on the filtered corpus. The distributions are shown in the right column of figure 2. We can observe that the overlap between the three classes is less important in this corpus than in the previous one. More specifically, the distributions of questions slightly move away from the two other classes (D, E).

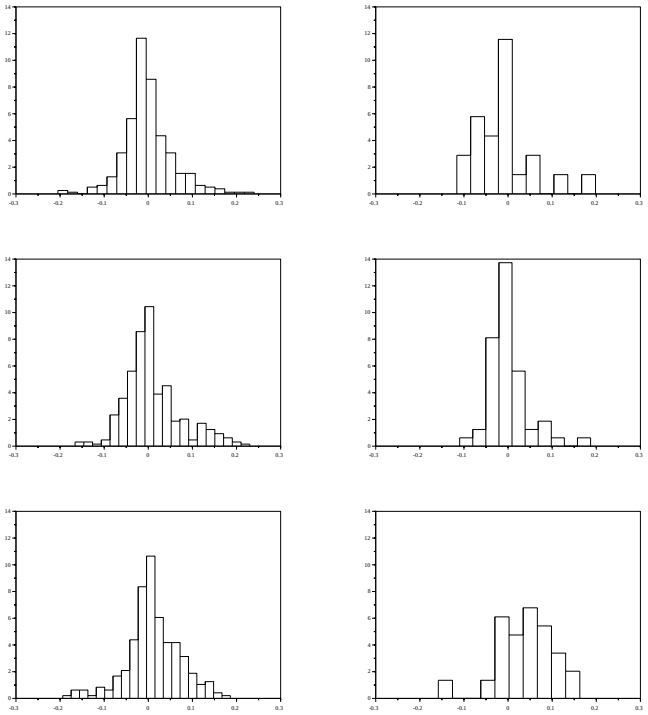

Fig. 2. Comparison of F0 slopes distributions on the original (left) and filtered (right) corpora. From top to bottom: exclamations, declarations and questions.

4.2. Automatic modality recognition with F0

The objective of this experiment is to evaluate the automatic recognition of sentences modality based on F0 only. We first build a training corpus from the original corpus according to the basic prosodic rules described in section 2: the linear regression on the four F0 values is computed as previously. Any "?" sentence with a regression slope greater than 0.03 is assumed to belong to class (Q̌). Similarly, any "!" sentence with a slope smaller than -0.03 is classified in the Ě class. Finally, any "." sentence with a slope between -0.01 and 0.01 is assumed to be declarative (Ď). The resulting *LR-filtered* training corpus is composed of 252 sentences (74 for Ď, 87 for Ě and 91 for Q̌). The test corpus is the test part of the manually labeled filtered corpus.

As we are mainly concerned by the dynamic evolution of F0 values, the feature vector is composed of the time derivatives of F0. The training database is then modeled by three Gaussian Mixture Models (GMM), one per modality.

Table 1 shows the confusion matrix for each class. We can observe that questions are better recognized than exclamatory and declarative sentences. This suggests that the final F0 slope is a more discriminative criterion between questions and other sentences than between declarative and exclamatory sentences. The global accuracy of this experiment is 54 %.

	Recognized class in [%]		
Pronounced class	Q̌	Ě	Ď
Q̌	**75**	11	14
Ě	18	**47**	35
Ď	16	40	**44**

Table 1. GMM's confusion matrix in %

4.3. Automatic modality recognition with F0 and energy

This second approach involves another prosodic attribute, the energy. Each sentence is represented by 20 features for F0 and 20 features for energy. Furthermore, we investigate the following semi-automatic method to build a new training corpus for the classifier: first, three GMMs are trained on the filtered corpus. Then, the second part of the original corpus, which has not been manually filtered, is recognized by these three GMMs to produce a new *GMM-filtered* training corpus.

Next, a multi-layer perceptron (MLP) is trained on this corpus. The MLP has three layers with 40 inputs, 25 neurons in the hidden layer and three outputs, according to sentences types. The test part of the filtered corpus is used as in the previous experiment. The resulting recognition accuracy is shown in table 2. The global accuracy of this experiment is 59 %, which outperforms the previous experiment. On the other hand, the recognition rate of questions alone has decreased. We may conclude that the new features are important to recognize declarative sentences. Conversely, they may introduce some confusion for interrogative sentences.

The global recognition rate is still not satisfactory. This may be due to either errors during manual labeling (see discussion in section 3), or lack of training data, or inadequacy of the prosodic features used for modeling. Indeed, it seems

	Recognized class in [%]		
Pronounced class	Q̌	Ě	Ď
Q̌	**53**	31	16
Ě	12	**47**	41
Ď	17	14	**69**

Table 2. MLP's confusion matrix in %

obvious now that other features than prosody (such as syntax and semantic) play a major role in sentence modality recognition.

4.4. Automatic modality recognition by a combination of both previous approaches

A conclusion of both previous experiments is that different features/classifiers better recognize different sentence modalities. We propose here to combine sequentially both classifiers as follows: First, the GMM classifier is applied to detect questions. The sentences recognized as questions are then definitely classified as Q̌ and removed from the test set. The remaining sentences are then recognized by the second module, which is composed of the MLP system described above. This linear combination has been chosen because of its simplicity, which makes it very easy to implement, and because of the better quality of the GMM to recognize questions. We believe that the recognition errors made by both recognizers (the GMM and the MLP) are partly complementary, and we have designed this experiment to test this hypothesis. Table 3 shows the confusion and recognition rate for each class.

	Recognized class in [%]		
Pronounced class	Q̌	Ě	Ď
Q̌	**84**	16	0
Ě	18	**41**	41
Ď	28	16	**56**

Table 3. Confusion matrix (in %) for the combined approach.

The global accuracy of this experiment is 61 %. We also note the good recognition rate, 84 %, of questions, which is very interesting for our application.

4.5. Questions recognition

In this experiment, we group together both D and E modalities, and we focus our efforts on detecting questions. This can be done in the statistical hypothesis testing framework, using likelihood ratios of class and anti-class models.

The GMM-filtered corpus is used for training question models and anti-models, hereafter called Q̌ and Q̄. The Q̌ model is a 2-mixture GMM, while the Q̄ model is a 6-mixture GMM. The test corpus is the same as before. Detection results are given in figure 3. We can observe that the results are less good than the ones obtained in the previous experiments, which is probably partly due to the GMM classifier.

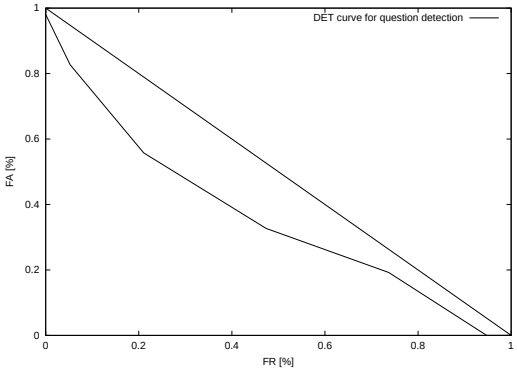

Fig. 3. DET curve for questions detection. False rejections are plotted on the X-axis and false acceptances on the Y-axis.

5. CONCLUSION

In this work, we compare two different sets of prosodic features and classifiers to recognize French sentences modality, in the context of an application that aims at enriching a talking head with such modal information. Experimental results show that 75 % of interrogative sentences, the most important type of modality in our application, can be recognized by using F0 features only, and about 70 % of declarative sentences with F0 and energy. Combining both classifiers increases questions recognition up to 84 %.

The tested systems give promising results, but need further improvements to be integrated into the target application. We identified two potential issues: insufficient prosodic features, and errors in the labeling of the corpus. The latter problem shall be solved by an upcoming project that will build a more suitable corpus for this task. The next step will consist of considering other prosodic features as well as non-prosodic clues, such as syntax and semantic.

6. ACKNOWLEDGEMENT

This work would not have been possible without the aid of Daniel Dechelot and Emanuel Didiot from French laboratory Loria, who is participed to the manual corpus relabeling.

7. REFERENCES

[1] R. O. Cornett, "Cued speech," in *American Annals of the Deaf*, 1967, vol. 112, pp. 3–13.

[2] P. Kral and J. Kleckova, "Speech recognition and animation of talking head," in *IWSSIP'03*, Prague, Czech Republic, September 2003.

[3] H. Gezundhajt, "La prosodie," in *http://www.linguistes.com/phonetique/prosodie.html*.

[4] V. Aubergé, "A gestalt morphology of prosody directed by functions: the example of a step by step model developed at ICP," in *1st Conf on Speech Prosody*, 2002, pp. 151–155.

[5] R. Kompe, *Prosody in Speech Understanding Systems*, Springer, July 1997.

[6] V. Strom, "Detection of accents, phrase boundaries and sentence modality in german with prosodic features," in *Eurospeech'95*, Madrid, 1995.

[7] J. Kleckova and V. Matousek, "Using prosodic characteristics in Czech dialog system," in *Interact'97*, 1997.

[8] K. Chongdok and Y. Hiyon, "Defining modality by terminal contours in standard korean," in *1st International Conference on Speech Sciences*, Seoul, 2002.

[9] H. Wright, M. Poesio, and S. Isard, "Using high level dialogue information for dialogue act recognition using prosodic features," in *ESCA Workshop on Prosody and Dialogue*, Eindhoven, Holland, September 1999.

[10] H. Wright, "Automatic utterance type detection using suprasegmental features," in *ICSLP'98*, Sydney, 1998, p. 1403.

[11] "http://www.recherche.gouv.fr/technolangue/," .

[12] A. de Cheveigne and H. Kawahara, "Comparative evaluation of F estimation algorithms," in *Eurospeech'2001*, Scandinavia, 2001.