Automatic Classification of French Spontaneous Oral Speech into Injunction and No-injunction Classes

Abdenour Hacine-Gharbi[1][a] and Philippe Ravier[2][b]

[1]LMSE Laboratory, University of Bordj Bou Arréridj, Elanasser, 34030 Bordj Bou Arréridj, Algeria
[2]PRISME Laboratory, University of Orleans, 12 Rue de Blois, 45067 Orleans, France

Keywords: Speech Injunction Classification, Massive Wild Oral Corpus, Feature Extraction, Wrapper Feature Selection, Prosodic Features, Static and Dynamic Features, GMM Models.

Abstract: The injunctive values are of particular interest for many studies dealing with oral speech interactions, e.g. in automatic meaning processing or in the field of language pathology understanding and therapy. We propose in this paper an automatic classification system using a subset of the RAVIOLI database in order to evaluate the role of prosody in the definition of injunctive values. RAVIOLI is constituted of more than 100 hours wild massive oral spontaneous speech. This work is a preliminary study that exploits a subset of 197 injunction values that have been labelled as exploitable utterances by two linguistic experts augmented by 198 of non-injunctive utterances. Many feature types were considered for this study: some classical features employed in speech community for automatic speech recognition tasks (LPCC, MFCC and PLP with their associated dynamic features) and some prosodic features (pitch and energy, with their associated dynamic features). The results clearly show the importance of prosodic features for the classification of the utterances into injunction or no-injunction classes and particularly the predominance of the log energy feature.

1 INTRODUCTION

The injunctive values are of particular interest for many studies dealing with oral speech interactions, e.g. in automatic meaning processing or in the field of language pathology understanding and therapy.

Injunction can be defined as enjoining somebody in doing something. In this utterance mode, the speaker expresses his willingness to obtain from the recipient a certain behaviour and aspire him to make the propositional content of the utterance act. In this context, imperative can be considered as the typical way to express the injunction (Nguyen Minh, Chinh, 2013) (Nguyen Minh, 2016). It can be used for any injunctive purpose, from the lightest (prayer, solicitation) to more peremptory (order, command). Imperative utterance is commonly described by syntactic indices (absence of subject) associated to morphological marks (some specific verbal form). It can be thus automatically detected because of this prototype belonging to a specific morphosyntactic schema. However, the imperative form does not express the injunctive value only and it can be used for other purposes such as the expression of the condition or question etc. As an example, the French 'vas-y' expression is used by some young French people in spontaneous interactions with the actual meaning 'oui'. This example shows a morphosyntactic schema of imperative form while not having any injunctive value at all. Conversely, some morphosyntactic schemas can be distant from the imperative structures like in the French injunction 'il faut partir maintenant, je vous demande de vous arrêter'. Injunctions can also be expressed in declarative or interrogative sentences like in 'Tu veux bien ouvrir la porte?'

So automatic classification of injunctions in oral corpus is not straightforward because it cannot solely rely on the numerous works dedicated to prosody classification. This study aims at achieving an automatic classification of injunctions in the RAVIOLI wild oral interaction data corpus (RAVIOLI means 'Reconnaissance Automatique des Valeurs Injonctives à l'Oral, Langue en Interaction'

[a] https://orcid.org/0000-0002-7045-4759
[b] https://orcid.org/0000-0002-0925-6905

Hacine-Gharbi, A. and Ravier, P.
Automatic Classification of French Spontaneous Oral Speech into Injunction and No-injunction Classes.
DOI: 10.5220/0009354406380644
In *Proceedings of the 9th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2020)*, pages 638-644
ISBN: 978-989-758-397-1; ISSN: 2184-4313

for Automatic Recognition of Injunctive Values in Interaction Oral Language). This corpus is constituted of authentic and massive speech oral French data gathered in a spontaneous noisy speech interaction environment with various speakers. In a first study, this work will be limited to a labelled subset of sentences containing the French word 'allez', which contain injunctive values. This work will give an answer to the role of prosody in the injunctive interpretation that emerges at the syntagm and sentence level.

More precisely, the principal first aim of this work is to evaluate the contribution of the prosodic descriptor for the task of speech injunction classification using Gaussian Mixtures Models (GMM) models. The study will investigate prosody descriptors (pitch, energy) and other typical descriptors used in speech processing, such that Linear Predictive Cepstral Coefficients (LPCC), Perceptual Linear Prediction coefficients (PLP) and Mel-Frequency Cepstrum Coefficients (MFCC) (Basu, Chakraborty, Bag, & Aftabuddin, 2017) (Wu, Falk, & Chan, 2011) (Hacine-Gharbi, Deriche, Ravier, Harba, & Mohamadi, 2013) with their first derivatives and second derivatives for dynamical modelling. Performance of the prosodic descriptors will be compared with performance of LPCC, PLP and MFCC descriptors.

The second aim of the paper is to select the relevant prosodic features for this task using a wrapper method.

The paper is organized as follows. Section 2 describes the injunctions classification system based on a brief state of the art of some prosodic classification systems. Section 3 introduces the RAVIOLI database, details the system and gives classification rate results for many experiments. In addition, results of a feature selection procedure are analysed before concluding the paper.

2 SPEECH INJUNCTIONS CLASSIFICATION SYSTEM

Several systems based on prosodic features have been proposed for pattern recognition tasks (Singh, Khan, & Pandey, 2012) (Mary & Yegnanarayana, 2008) (Ferrer, et al., 2015) (Szaszak, Tündik, & Gerazov, 2018). In (Hacine-Gharbi, Petit, Ravier, & Nemo, 2015), the authors have proposed an automatic system for the classification of the speech signal into semantic categories labelled "conviction (CV)" and "lack of conviction (NCV)" in the use of the French single word 'oui'. Each one of these semantic categories has been associated to a class in an automatic classification task, using hidden Markov models (HMM) for class modelling. The authors have more specifically studied the contribution of the prosodic features in the discrimination between the two classes. These features are the pitch and the energy as well as their dynamic features delta and delta-delta. In (Hacine-Gharbi, A.; Ravier, P.; Nemo, F., 2017), the authors go one step further with the extraction of the local and global prosodic features, combined with SVM classifier to classify the signal of French 'oui' into classes CV or NCV. In (Hacine-Gharbi, A.; Ravier, P., 2019), the authors have investigated the combination of MFCC coefficients with prosodic features for emotions recognition based on modelling each emotion class by GMM.

The purpose of the present system is to classify speech audio signals into injunctive class (INJ) or non-injunctive class (NINJ). We propose the use of GMM combined with feature extraction method for this task of classification. In this paper, we will consider prosodic features and classical spectral features.

Such a system requires a learning phase in order to obtain a GMM model for each class followed by a testing phase in order to identify the class an unknown utterance signal belongs to. Thus, the dataset is divided into a training dataset for the learning phase and a testing dataset useful for evaluating performance of the classification system. Each phase requires the extraction of features that contain information useful for discriminating the two classes INJ and NINJ.

Figure 1 illustrates the diagram of our automatic classification system of speech occurrences into injunctive and non-injunctive classes.

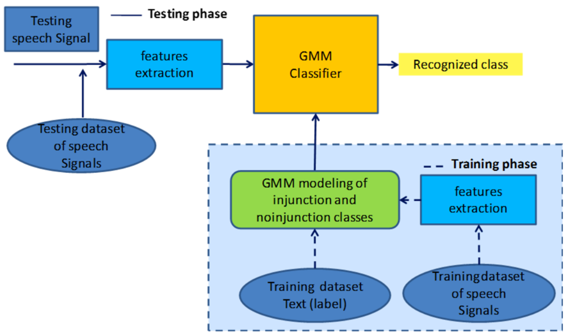
Testing speech Signal
Testing phase
features extraction
GMM Classifier
Recognized class
Testing dataset of speech Signals
Training phase
GMM modeling of injunction and noinjunction classes
features extraction
Training dataset Text (label)
Trainingdataset of speech Signals

Figure 1: Diagram of the GMM classification system in the two classes INJ or NINJ. The system is composed of the training phase (dashed lines) which learns the GMM injunction models using the occurrences of the training dataset with their corresponding text; the testing phase (full lines) decides whether the test signal is injunctive or not.

Implementation of the system is carried out using HTK tools (Hidden Markov Model Toolkit (Young, Kershaw, Odell, & Ollason, 1999)), in which we consider GMM as HMM model with one state modelled by GMM of nG Gaussian components with diagonal covariance matrix (Hacine-Gharbi, A.; Ravier, P., 2019).

In order to take into account dynamic evolutions of the data, we used the pitch and the energy prosodic features with their respective first and second derivatives, which forms 6-components feature vectors. The derivatives are dynamic features that quantify speed (derivative D) and acceleration (second derivative A) in static features changes between consecutive frames.

These prosodic features will be compared to the classical spectral features (LPCC, PLP and MFCC) augmented by their dynamics features in terms of classification performance and number of features.

In the next section, we will describe the system with the RAVIOLI database and we will test different configurations of the system with different feature vector compositions. We will discuss the relevance of the components of the vectors for the injunction classification system.

3 EXPERIMENTS AND RESULTS

3.1 Description of the RAVIOLI Database

The system performance evaluation requires a database of audio signals that belong to injunctive and non-injunctive classes. In RAVIOLI project, a database of injunctive and non-injunctive utterances was collected from several environments and characterized by a great variability and complexity of spontaneous speech signals (age, gender, speaker, noise, emotions, ...). One of the objective of the RAVIOLI project is to precisely determine the role of prosody in the interpretation of injunctive values.

The database is therefore constituted of injunctive utterances produced in authentic oral interactions collected from the ESLO2 database (http://eslo.huma-num.fr/).

The original RAVIOLI database contains about 106 hours of recording distributed in 150 sound files with as much as transcription files (obtained using the transcriber software). The RAVIOLI dataset is constituted of the four following modules:

- School: a pupils classroom (71h00)
- 24h: one day follow-up of a person (9h44)
- Itinerary: city itinerary asking and questioning (5h42)
- Meal: family or friends mealtimes (18h50)

A fifth module of interviews were removed because of a lack of injunctive values.

3.2 The Injunction Classification System

3.2.1 Dataset Repartition

The RAVIOLI speech database have originally been recorded at the 44100 Hz sampling frequency. Records have been down-sampled to 16000 Hz for reducing the computation cost. We selected a dataset of injunctions and non-injunctions as a part of the database for this preliminary work.

The selected dataset was divided into a training dataset composed of 100 injunctive utterances and 99 of non-injunctive utterances and a testing dataset composed of 97 injunctive utterances and 99 non-injunctive utterances. Each injunctive utterance is constituted of spontaneous French speech including the French word “aller” (or “allez”), which is used as key word for classifying the utterances by linguistic experts in RAVIOLI project. The use of the key word is not necessarily associated to an imperative form. The non-injunctive utterances are composed of many different speech objects like sentences, single words, interjections, murmurs, noise …

We also recall that both classes INJ and NINJ are composed of utterances selected by linguistic experts in wild oral interactions. Therefore, the dataset is constituted of speech recorded for different speakers in different places and different conditions. The recordings include the speech of a professor or pupils in a noisy classroom; the oral interactions at mealtimes; the speech of a person giving direction explanations to another one asking for his way in the street; oral exchanges during work sessions. Notice that only the fully exploitable utterances were considered for the constitution of the INJ dataset in this experiment (that is the utterances considered as clearly showing injunctive values with a common decision given by at least two linguistic experts).

The segmented utterances may have a length of 0.2 second to 5 seconds.

3.2.2 Configuration of the System

Each utterance signal (either INJ or NINJ) is then converted into a sequence of features vectors computed each 10 ms on 30 ms analysing Hamming windows, using the 'Hcopy' command of HTK library. The sequences of the training dataset are used to model each class labelled INJ or NINJ by a GMM model using the command 'HEREST'. Next, each sequence of vectors of the testing dataset is classified using the command 'HVITE'. Finally, performance evaluation is done by using the command 'HResult'.

The quality of the classification system is evaluated by a classification rate CR defined as:

$$\mathrm{CR} = \frac{O - M}{O} \qquad (1)$$

where O is the total number of occurrences given at the input of the classifier and M is the number of misclassified occurrences.

3.2.3 Feature Extraction

Many feature types are investigated. A set of 12 features are calculated, for each type LPCC, MFCC and PLP. This set of 12 static features is augmented by the 'D' and 'A' dynamic features, which produces 36-components feature vectors. The prosodic features are the log energy 'E' values extracted from HTK software and the pitch 'PI' values that are computed each 10 ms by Praat software (Boersma & Weenink, 2014). The 'D' and 'A' dynamic features are also calculated on these prosodic features using 'Hcopy' command of HTK Tools library.

Finally, sequences of 114-components feature vectors are obtained by following the computation procedure depicted in Figure 2.

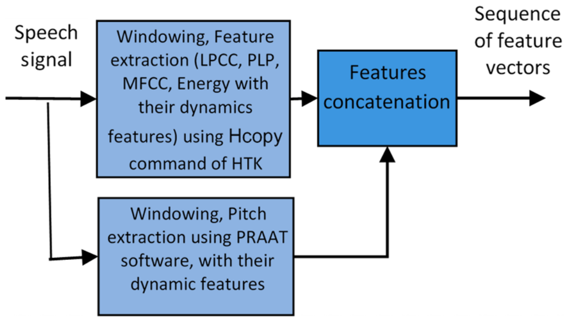
Speech signal
Windowing, Feature extraction (LPCC, PLP, MFCC, Energy with their dynamics features) using Hcopy command of HTK
Features concatenation
Sequence of feature vectors
Windowing, Pitch extraction using PRAAT software, with their dynamic features

Figure 2: Feature extraction scheme for the classical and prosodic types.

3.3 Classification Statistics

In this section, the conducted experiments aim at practically demonstrate the importance of the investigated feature types in the classification results. A comparative study is performed between the feature types, the speech classical ones and the prosodic ones, the static and dynamic ones. The different configurations of descriptors are noted TYPE_EDA in which TYPE is LPCC, MFCC or PLP, 'E' represents the energy (actually the log energy), 'D' is the derivative Δ (speed) and 'A' is the double derivative ΔΔ (acceleration). In these configurations, the letters are optional. The TYPE can also be a single prosodic feature, like 'PI' for the pitch or 'E' for the log energy. For each configuration, the conducted experiment identified the optimal Gaussian components number nG which gives the highest CR value when changing the number of Gaussians in the set {1,2,4,8,16,32,64,128,256}.

Finally, the results shown in Table 1 give the CR values for each configuration with the optimal Gaussian components number.

Table 1: Classification rate **CR** and optimal Gaussian number **nG** of GMM models as a function of the descriptor type configurations. The dimension **d** of each descriptor type is added.

	MFCC _EDA	**MFCC _DA**	**LPCC _EDA**	**LPCC _DA**	**PLP _EDA**	**PLP _DA**
d	**39**	**36**	**39**	**36**	**39**	**36**
CR	52.55	52.04	51.53	54.59	51.53	48.47
nG	4	4	16	2	8	4

	PI	**PI_DA**	**E**	**E_DA**	**PI, E**	**PI_DA, E_DA**
d	**1**	**3**	**1**	**3**	**2**	**6**
CR	68.37	64.80	85.20	83.67	72.45	68.37
nG	16	8	4	8	64	128

The use of the entire set of 114 features also gives CR=53.06% with nG = 2.

This table clearly shows the importance of the energy prosodic feature all alone, either used in its static or dynamic version. The best result is obtained with a reduced number of 4 Gaussians and the classification rate reaches 85.20% with only one feature, the energy 'E'. When combined with pitch feature, performance results decrease. The use of non-prosodic features give very bad results (around 50%). Note that these classical features are known to be appropriate for automatic speech recognition tasks, for example the classification of key word utterances. Even if the word 'allez' can be heard in the utterances of the INJ dataset, this characteristic is not sufficient for discriminating the utterances of the NINJ dataset that excludes this key word. This may be due to the too noisy conditions and the fact that the keyword is embedded in the utterances, which makes the classification more difficult. This also means that the prosodic features are good discriminators for the classification of utterances into injunction or no-injunction classes (even if the injunctive utterances do not all show imperative forms).

Furthermore, adding energy prosodic feature 'E' to the classical features do not improve the results because information brought by this new feature is drown in a too high number of non-information features. In order to deeply analyse the role of prosodic features in our classification problem, we thus investigate feature selection method on the limited prosodic number of features using the 'wrapper' method.

3.4 Feature Selection using Wrapper Method

The two principal categories of feature selection methods are the 'filter' methods and the 'wrapper' methods described in (Kohavi & John, 1997). The 'filter' methods evaluate the relevance of features for describing the classes based on the quantity of information the features bring for explaining the classes. This strategy is useful when a high dimension feature vector needs to be considered for classification task, which is not the case regarding results of the previous section. Contrary to the 'filter' methods, the 'wrapper' methods are based on the optimality criterion of highest classification rate, which however depends on the classifier to be built (Giannoulis & Potamianos, 2012). In this work, we applied the wrapper-based sequential forward search (SFS) algorithm (Kohavi & John, 1997) considering a set of the six prosodic features alone for the selection. Indeed, the previous section took away the numerous classical features, which makes the computational cost for this reduced size tractable. The SFS algorithm sequentially adds at each selection step the feature that gives the highest CR. This algorithm has been used in (Hacine-Gharbi, Petit, Ravier, & Nemo, 2015) (Nait Meziane, et al., 2017) (Hacine-Gharbi, A.; Ravier, P., 2018).

Table 2: Prosodic feature selection using 'wrapper' SFS algorithm for two number of Gaussians configurations nG=4 and nG=8 of GMM models. The #i number refers to the ith selected feature in the forward selection process.

nG=4	**#1**	**#2**	**#3**	**#4**	**#5**	**#6**
TYPE	E	E_A	E_D	PI_A	PI	PI_D
CR	85.20	84.18	81.63	78.06	75.00	58.67

nG=8	**#1**	**#2**	**#3**	**#4**	**#5**	**#6**
TYPE	E	E_D	PI_D	PI_A	E_A	PI
CR	85.20	82.14	86.22	78.06	81.63	67.86

As the features selection step depends on CR results, we also propose to find the relevant subset of features for each gaussian number.

Table 2 shows the order of the selected features using either nG=4 or nG=8 for the Gaussian number of GMM models.

This table confirms the predominance of the 'E' energy feature, which is always the first selected feature. The maximum CR=86.22% is obtained with the combination of three features {E, E_D, PI_D} and nG = 8. Augmenting the feature set by the dynamics of the pitch values (PI_D) improves prosody evaluation, which helps discriminating injunctive utterances from non-injunctive ones. Note that this result regarding prosody was already observed in (Hacine-Gharbi, Petit, Ravier, & Nemo, 2015) which results highlight that the prosodic features {PI_D, E, PI} provide better performance than the set of other features for the task of classification of the uses of the French word 'oui' as convinced or unconvinced classes in the speaker independent mode. Also, related to the current injunction study, this result regarding prosody is confirmed by the study of (Basirat, Patin, & Moreau, 2018) who found that when perception and production of speech is impaired, intonation differences are higher in control subjects than impaired ones (Parkinson's disease (PD) subjects). The study evaluated the intonation difference between yes-no questions and statements in French of France. It resulted that PDs marked their questions and statements less than controls. Particularly, the authors observed a smaller pitch rise in PDs compared to controls. Therefore, impairment affects prosody by lowering the prosodic dynamics and thus the meaning processing. Conversely, high prosodic dynamics contribute to meaning processing thus helping injunctive values interpretation.

Table 3: Prosodic feature selection using 'wrapper' SFS algorithm by exploring different Gaussian numbers in the set {1,2,4,8,16,32,64,128,256}. The reported values give the best CR and the selected features set for each nG value.

#nG	**1**	**2**	**4**	**8**
YPES	{PI}	{E}	{E}	{E,E_D,PI_D}
CR	63.78	81.63	85.20	86.22

#nG	**16**	**32**	**64**	**128**	**256**
YPES	{E}	{E}	{E}	{E}	{E}
CR	84.69	84.18	83.67	84.18	84.18

In order to confirm results of Table 2, we also investigated the influence of the Gaussian number values. This number was changed in the set {1,2,4,8,16,32,64,128,256} and for each tested nG value we evaluated the best CR obtained in the SFS selection process as well as the selected features set. The results reported in Table 3 confirm the importance of the energy feature 'E' as single feature to be selected in all the cases except nG = {1,8}. The case nG = 8 gives the best results with three features.

We have also investigated HMM modelling with many states in order to examine whether if a richer modelling improves the classification performance results. Table 4 gives the CR results by increasing the number of HMM states by taking nG = 8 which was the best configuration of the previous experiments.

Table 4: Evolution of the CR with respect to the number of states nS considered for HMM modelling. The number of Gaussians is set to nG = 8.

nS	1	2	3	4	5	6	7	8
CR	86.22	68.88	62.24	73.47	61.73	71.43	70.92	69.39

The results show that GMM modelling is preferable to HMM modelling with many states that are not appropriate for this classification task.

4 CONCLUSIONS

This study firstly aimed at evaluating whether if an automatic system can be able to classify some utterances from wild data corpus of spontaneous oral speech into the injunction or no-injunction classes. The proposed built system achieved the best classification rate of 86.22% on a subset of 199 training utterances and 196 testing utterances using a total of 395 utterances of the RAVIOLI database (197 INJ utterances and 198 NINJ utterances). The system was a GMM classifier composed of 8 Gaussians using three prosodic features.

This study secondly aimed at evaluating which type of features can be appropriate for this classification task. The results showed the predominance of the log energy feature. A slight improvement was obtained using other dynamic prosodic features of energy and pitch. Hence, the prosodic features are relevant for the task of classification into injunction and no-injunction classes. Importantly, the spectral features classically used for automatic speech recognition tasks are not appropriate at all.

The main aim of the RAVIOLI project is the characterization and interpretation of injunctive values. This project objective constitutes the natural extension of this work by investigating a larger dataset of the RAVIOLI database and segmenting injunction utterances in the continuous spontaneous speech. Other classifiers like SVM, KNN or ANN can be investigated in order to validate the relevance of the prosodic features. Other tasks will be the automatic clustering of prosodic features because this preliminary work have shown the importance of this type of feature for injunctions characterisation.

ACKNOWLEDGEMENTS

This work is funded by the région Centre Val-de Loire, France, under the contract APR IA Ravioli #17055LLL (lotfi.abouda@univ-orleans.fr, leader). This collaborative work implies three laboratories of University of Orléans (LLL, PRISME/IRAuS, LIFO) and one laboratory from University of Tours (LIFAT). All the persons involved in this project are acknowledged for their participation.

REFERENCES

Basirat, A., Patin, C., & Moreau, C. (2018). Relationship between Perception and Production of Intonation of French in Parkinson's Disease. *9th International Conference on Speech Prosody*, (pp. 809-813). Poznan, Poland.

Basu, S., Chakraborty, J., Bag, A., & Aftabuddin, M. (2017). speech, A review on emotion recognition using. *International Conference on Inventive Communication and Computational Technologies (ICICCT)* , (pp. 109-114).

Boersma, P., & Weenink, D. (2014). *Praat: doing phonetics by computer. [Computer program]. Version 5.3.75, www.praat.org.*

Ferrer, L., Bratt, H., Richey, C., Franco, H., Abrash, V., & Precoda, K. (2015). Classification of lexical stress using spectral and prosodic features for computer-assisted language learning systems. *Speech Communication*, 31-45.

Giannoulis, P., & Potamianos, G. (2012). A Hierarchical Approach with Feature Selection for Emotion Recognition from Speech. *the 8th International Conference on Language Resources and Evaluation (LREC'12)*, (pp. 1203–1206). Istanbul, Turkey.

Hacine-Gharbi, A., Deriche, M., Ravier, P., Harba, R., & Mohamadi, T. (2013). A new histogram-based estimation technique of entropy and mutual information using mean squared error minimization. *Computers and Electrical Engineering, 39*(3), 918-933.

Hacine-Gharbi, A., Petit, M., Ravier, P., & Nemo, F. (2015). Prosody based Automatic Classification of the Uses of French 'Oui' as Convinced or Unconvinced Uses:. *Proceedings of the 4th International Conference on Pattern Recognition Applications and Methods (ICPRAM)*, (pp. 349-354). Lisbon, Portugal.

Hacine-Gharbi, A.; Ravier, P. (2018). Wavelet Cepstral Coefficients for Electrical Appliances Identification using Hidden Markov Models. *Proceedings of the 7th International Conference on Pattern Recognition Applications and Methods (ICPRAM)*, (pp. 541-549). Madeira, Portugal.

Hacine-Gharbi, A.; Ravier, P. (2019). On the optimal number estimation of selected features using joint histogram based mutual information for speech emotion recognition. *Journal of King Saud University - Computer and Information Sciences*, in press.

Hacine-Gharbi, A.; Ravier, P.; Nemo, F. (2017). Local and global feature selection for prosodic classification of the word's uses. *Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods (ICPRAM)*, (pp. 711-717). Porto, Portugal.

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. *Artificial Intelligence, 97*(1-2), 273-324.

Mary, L., & Yegnanarayana, G. (2008). Extraction and representation of prosodic features for language and speaker recognition. *Speech Communication*, 782-796.

Nait Meziane, M., Hacine-Gharbi, A., Ravier, P., Lamarque, G., Le Bunetel, J.-C., & Raingeaud, Y. (2017). Electrical Appliances Identification and Clustering using Novel Turn-on Transient Features. *Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods (ICPRAM)*, (pp. 647-652). Porto, Portugal.

Nguyen Minh, C. (2016). L'impératif en français parlé. *5ème Congrès Mondial de Linguistique Française* (p. 15). Tours: EDP Sciences.

Nguyen Minh, Chinh. (2013). *L'injonction dans le Français parlé d'une approche en langue à une analyse de corpus.* University of Paris 3: PhD thesis.

Singh, N., Khan, R. A., & Pandey, R. (2012). MFCC and Prosodic Feature Extraction Techniques: A Comparative Study. *International Journal of Computer Applications* , 9-13.

Szaszak, G., Tündik, M. A., & Gerazov, B. (2018). Prosodic stress detection for fixed stress languages using formal atom decomposition and a statistical hidden Markov hybrid. *Speech Communication*, 14-26.

Wu, S., Falk, T. H., & Chan, W.-Y. (2011). Automatic speech emotion recognition using modulation spectral features. *Speech Communication, 53*, 768–785.

Young, S., Kershaw, D., Odell, J., & Ollason, D. (1999). *The HTK Book.* Cambridge: Entropic Ltd.