Prosody based Automatic Classification of the Uses of French 'Oui' as Convinced or Unconvinced Uses

Abdenour Hacine-Gharbi[1,3], Mélanie Petit[2], Philippe Ravier[1] and François Némo[2]
[1]PRISME laboratory, University of Orléans, 12 rue de Blois BP 7744 – 45067 Orléans, France
[2]Laboratoire Ligérien de Linguistique, University of Orléans, Orléans, France
[3]LMSE, University of Bordj Bou Arréridj, Bordj Bou Arréridj, Algéria

Keywords: Intonation Classification, HMM, Prosodic Features, Reduction Dimensionality, Curse of Dimensionality, Wrappers Feature Selection, Categorization of Word's Uses.

Abstract: When working with oral speech, the issue of natural meaning processing can be improved using easily available prosodic information. Only recently, semanticists have started to consider that the prosodic features could play a key role in the interpretation and classification of different word's uses. In this work, we propose a prosodic based automatic system that allows to classify the French word 'oui' into one of the classes 'conviction' or 'lack of conviction'. To that aim, a questionnaire inspired from opinion polls has been created and permitted to obtain 118 occurrences for both classes of 'oui'. Combined with feature selection procedure, the best classification rates decreases from 85.45% (speaker dependent mode) to 79.25% (speaker independent mode which is closer to an application). Interestingly, we also introduce the 'shuttle' principle that seeks to validate the semantic interpretation thanks to prosodic analysis.

1 INTRODUCTION

Linguists have always been confronted with the fact that words have many different meanings (polysemy). Recently, with authentic oral spoken data becoming available in large quantities, they have been confronted with the reality that because this diversity appears to be much wider than previously thought, the semantic description, categorization and classification of these word's uses required to be greatly improved as far as prosody was concerned.

For the past 10 years, in order to do so, semanticists have started to consider that the prosodic features could play a key role in the interpretation and classification of different word's uses. The first doctoral works (Petit, 2009) entirely dedicated to this issue have shown that the prosodic features could serve as an explanation of aspects of the word's interpretation and as a key to the discrimination of the different word's uses.

Indeed, the prosody or intonation is an important information source of spoken communication. It is the reason why prosody plays a significant role in syntactic, semantic and pragmatic interpretation (Kompe, 1997) (Rosenberg, 2009) (Szaszák, Sztahó, & Vicsi, 2009). Moreover, several studies have shown the advantages of the prosody in many spoken language processing tasks including: automatic speech recognition (Hasegawa-Johnson & all, 2004, Chao Wang, 2001), speaker identification (Manganaro, Peskin, & Shriberg, 2002), language recognition (Mary & Yegnanaray ana, 2008), Automatic Age Estimation (Spiegl & all, 2009), Automatic Classification of Dialog Acts (Shriberg & all), emotions states (Juslin & Laukka, 2003).

In the semantics field, it has been proven that study of the prosodic pattern of what was believed to be a single use (interpretation-type) of a sign was actually revealing the existence of various use-types. These use-types could be classified according to their prosodic pattern which means that the study of these patterns could allow for much more precise semantic descriptions.

Refining the semantic description of word's uses can be of great interest in spoken langage interpretation. For example, the same word frequently occuring in an oral opinion poll may have different meanings and interpretations depending on its detected prosodic pattern. Because of intrinsinc large databases related to this industry, achieving such a goal necessitates both automtaic detection of

Hacine-Gharbi A., Petit M., Ravier P. and Némo F..
Prosody based Automatic Classification of the Uses of French 'Oui' as Convinced or Unconvinced Uses.
DOI: 10.5220/0005293103490354
In *Proceedings of the International Conference on Pattern Recognition Applications and Methods* (ICPRAM-2015), pages 349-354
ISBN: 978-989-758-077-2

some specific words within the answers and the prosodic based automatic classification into one of its individual categorized identified uses.

Moreover, as for the study of the prosodic pattern themselves, it has been shown that a large data bank of all the uses of a given sign was necessary. Up-scaling the standard size of a few hundred into thousand implies automatic processing and automatic classification.

Centered on the uses of French oui and English yes, the DIASEMIE project has thus started to build such data bases and to categorize individual uses. What is presented here is one aspect of this process consisting in the identification of semantic features whose presence or absence in a given use can be tested. The feature at stake has been labeled "conviction" and "lack of conviction" and is associated with specific contexts of use of French "oui". Each one of these semantic features will be associated to a class in an automatic classification task.

The purpose of this paper is to investigate the task of automatic prosodic classification. The final goal is the categorization of word's uses that will be achieved by a systematic comparison of the semantic and prosodic characterization of the uses.

After a description of the classification system in the section 2, experiments and results will be presented in section 3 with a discussion on the systematic comparison of the characterization of the uses.

2 INTONATION CLASSIFICATION OF THE WORD 'OUI'

A wide range of machine learning techniques have been applied to the problem of automatic intonation classification. In (Szaszák, et al., 2009), a prosodic hidden Markov model (HMM) based modality recognizer has been developed. In (Shriberg & Stolcke, 2004), the authors have described a direct modeling approach of prosody in various speech technology tasks using either Gaussian mixture model (GMM) or decision trees. In (Fernandez & W. Picard, 2002), the authors have studied the use of Support Vector Machines (SVM) and discriminative learning techniques on the task of automatic classification of dialogue acts (DAs) from prosodic cues.

These methods can be grouped into two categories. The first category considers the prosodic pattern as a sequence of observation vectors. The components of the observation vectors are prosodic parameters computed during an analysis step. The sequences of vectors are used for training a Hidden Markov Model, each sequence being associated to a class.

Classifying consists in deciding whether any unknown computed prosodic sequence belongs to one class or to another.

The second category considers the prosodic pattern as a vector of statistical values of prosodic parameters (mean, standard deviation, …).

Using these prosodic statistical vectors, a model for each class can be trained such as GMMs, SVM models or artificial neural network models.

2.1 Set-up Procedure of the HMM based Recognition System

In the present work, the prosodic pattern is a sequence of prosodic vectors belonging to a class of convinced or unconvinced uses which can be represented by a HMM. Figure (1) presents the outline of our automatic classification system of prosodic patterns into word's use.

The implementation of this system is mainly based on the HTK library (Hidden Markov Model Toolkit) (Young, et al., 1999).

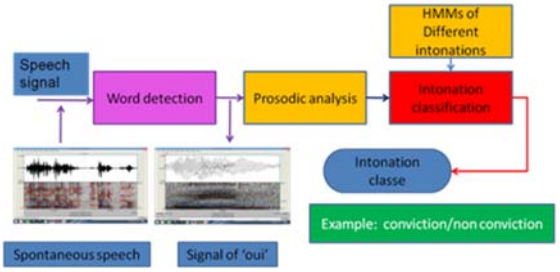

Figure 1: automatic prosody based classification system into word's use.

The classes of interest are labeled as 'conviction' and 'lack of conviction' expressed in the word 'oui'. Notice that the described system can be generalized for any other word, *e.g.* 'enfin' or 'voila'.

A supervised automatic classification system needs the achievement of two steps: the first one consists in detecting the word 'oui', this step has been manually achieved and will not be discussed in the following; the second step consists in labeling each occurrence of the word 'oui' of the database into the class 'conviction' and 'lack of conviction', this pre-processing step will be discussed in section 3.1.

A HMM based classification system can be decomposed into two classical phases, the training phase and the testing phase. The database is therefore split into a training database and a testing database. Both phases rely on a prosodic analyzing step which consists in transforming the temporal signal of word 'oui' into a sequence of vectors which components are prosodic features. The description of the prosodic features is given in the next section.

During the training phase, the system learns occurrences of the training database. The result is HMMs that represent classes through their prosodic vector sequences ('using HRest' command of the HTK library).

During the testing phase, the sequence of prosodic vectors of an unknown occurrence is proposed to the classifier ('using Hvite'). The classification decision is made taking the highest probability between the two classes.

This classification system has been experimented on a self made database constituted of a relatively small number of occurrences. Thus, the validation of such a system could face the curse of dimensionality problem (a performance decrease with an increase of the number of prosodic components) (Jain, et al., 2000). We therefore propose a feature selection step that will be introduced in section 2.3.

2.2 Definition of the Prosodic Feature Vector

Typical features that characterize prosody can be the pitch f_0 (Hz) and the energy E (dB). Thanks to PRAAT software (Boersma & Weenink, 2014), these parameters are computed every 10 ms on 30 ms analyzing windows of the temporal signal corresponding to an occurrence of the word 'oui'. A dynamic description of these static parameters f_0 and E is added by computing differential parameters of first and second order Δ and $\Delta\Delta$ using HTK library. Thus, each occurrence of the word 'oui' is represented by a sequence of vectors with 6 prosodic components noted as E, f_0, ΔE , Δf_0 , $\Delta\Delta E$, $\Delta\Delta f_0$.

The issue of the HMM based classifier is to make a decision for assigning the use of the word 'oui' in a predefined class, from a sequence of vectors composed of 6 prosodic parameters.

In our application, the HMM structures associated to the classes 'conviction' and 'lack of conviction' are composed of 5 states per class (and one state more for input and output) with mixture of 3 Gaussians per state.

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

$$TC = \frac{N - S}{N}$$

where N is the total number of occurrences given at the input of the classifier and S is the number of misclassified occurrences.

2.3 The Feature Selection Problem

Dimensionality reduction of the feature vectors can be achieved using features selection algorithms which select a subset of relevant feature from an initial set of features. These algorithms can be grouped into approaches that are classifier dependent ('wrapper' methods) and classifier independent ('filter' methods). Despite the wrapper methods have the disadvantage of a considerable computational expense; they have higher learning capacity in terms of over fitting (Brown, et al., 2012). So, in our work, we adopt wrapper methods because the disadvantage is minimized using only 6 features as initial set in the features selection. In particular, we use a forward sequential algorithm in which we select one feature at each step of selection. Moreover, the small size of the database makes the algorithm computationally tractable.

3 EXPERIMENTS AND RESULTS

3.1 Database Elaboration

In order to test the feasability of categorization of word's uses based on prosodic features, a small oral corpus has been created inspired from questions that can be asked in real opinion polls. The motivation for the construction of this database was to rapidly collect many instances of the word 'oui' thanks to questions leading to pronouncing the word 'oui' with expression of 'conviction' or 'lack of conviction'. The questionnaire is composed of 4 series with 10 questions each. Each series tackles more and more polemic topics (personal phone use, sport, European Union and politics). A group of 8 women and 17 men, all French native speakers, answered to this questionary. They were fully informed about the experimental procedures and all gave their signed consent.

It was difficult to label all the occurences of the word 'oui' in the dichotomy 'conviction' and 'lack of conviction', either because the conviction issue was not at stake, or because the word 'oui' expressed another feeling (pride, lassitude…). A

total of 52 occurrences for the class 'lack of convition' was obtained (taking into account some duplications of 'oui') and 66 occurrences were obtained for the class 'conviction'. The semantic categorization has been perceptually achieved using co-textual criteria.

3.2 Experimental Results of Classification

In the first experiment, we consider an intonation classification into the uses of the word 'oui' in speaker dependent mode. The database has been split into a set of 53.39% of the occurrences for the training phase and 46.61% for the testing phase. In this mode, the speakers participating in the testing phase have already participated in the training phase. In the second experiment, we consider an intonation classification system which is speaker independent. In this mode, the database has been split into a set of 55.08% of the occurrences for the training phase and 44.92% for the testing phase. The database division slightly differs with respect to the speaker dependent case because of the constraint of balanced occurrences number between the phases. The two modes considered allow us to quantify the influence of the speaker identity on the system performances.

The results show a classification rate of 80% in speaker dependent mode with the use of 6 features defined below. In the second mode, the results show a classification rate of 66.04% which demonstrates performance destruction compared to the first mode. This can be justified by an important prosodic variability (example: pitch variability) in the second mode caused by speakers inter-variability.

However, the relative reduced size of the database let the question of features relevance arise.

In order to give an answer to this question, we thus propose to add a feature selection step. This issue is discussed in the next section.

3.3 Feature Selection

We propose to select the most pertinent features for the discrimination between the two classes 'conviction / lack of conviction' in the 'oui' database. The wrapper algorithm with the forward strategy has been applied in the first and second classification modes

The different steps of this algorithm are:

1. F= {E, f_0, ΔE , Δf_0 , ΔΔE, $\Delta\Delta f_0$} , S={}, n=6 (initial number of parameters),
2. - Evaluate the classification rate (CR) for each feature $p_i \in F$.
   - Select the first feature $p_{\pi 1}$ with:
   $p_{\pi 1} = arg\max_{p_i \in F} CR(p_i)$,
   - F=F-$\{p_{\pi 1}\}$, S=$\{p_{\pi 1}\}$,
3. - In the iteration j, for each $p_i \in F$, evaluate the classification rate CR using the following set of features: $S \cup \{p_i\}$,
   - Select the feature $p_{\pi j}$ with:
   $$p_{\pi j} = arg\max_{p_i \in F} CR(S \cup \{p_i\})$$
   - F=F-$\{p_{\pi j}\}$, $S = S \cup \{p_{\pi j}\}$),
4. Repeat the step 3 until j=n,
5. Give the output set S that yields the maximum CR.

The table I displays the classification rate (CR) as a function of the number of selected features j in the speaker dependent mode.

Table I: Classification rate (CR) as a function of the number of selected features in the speaker dependent mode.

j	*1*	*2*	*3*	*4*	*5*	*6*
Feature selected at Iteration j	*E*	*Δ f_0*	*ΔΔ f_0*	*ΔE*	*ΔΔE*	*f_0*
CR (%)	72.73	83.64	*81.82*	***85.45***	81.82	80.00

From Table I, a number of remarks can be made:

- The set of features {E, Δf_0, ΔΔf0, ΔE} provides better performance (85.45%) than the set of all features (80%). This can be explained by the curse of dimensionality phenomenon caused by the lack of data for modelling the classes with a set of 6 prosodic features.
- The dynamic prosodic feature *Δf_0* with the energy play an important role for this task of classification in the speaker dependent mode.
- The statistic feature pitch *f_0* is not relevant for this task of classification.

The table II displays the classification rate (CR) as a function of the number of selected features j in the speaker independent mode.

Table II: Classification rate (CR) as function of the number of selected features in the speaker independent mode.

j	*1*	*2*	*3*	*4*	*5*	*6*
Feature selected at Iteration j	*Δ f_0*	*E*	*f_0*	*ΔΔ f_0*	*ΔE*	*ΔΔE*
CR (%)	*71.70*	*73.58*	***79.25***	*75.47*	*71.70*	*66.04*

From Table II, a number of remarks can be made:

- The set of features {*Δf_0, E, f_0*} provides better

performances (79.25%) than the set of all features (66.04%).

- The dynamic prosodic feature Δf_0 and the energy also play an important role for this task of classification.

3.4 Discussion

Previous work in the area has shown that a crucial part of the word's uses categorization consists in a systematic comparison of the semantic and prosodic characterization of the uses. On the one hand, all uses with the same prosodic patterns are each other compared. On the other hand, all uses with the same semantic categorization are compared. If semantic characterization is always reliable (what is said is true), it cannot be robust (i.e. it does not tell all the truth) because the characterization may change with co-textual criteria. The categorization has therefore to be improved through what we call the "shuttle" process.

As we shall see, it follows from this that the success of automatic classification and clustering cannot be measured only by its capacity to predict the initial categorization, but by its capacity to "fail rightly" whenever apparent "errors" of classification are proving that the semantics of the uses at stake is more complex than initially understood, for example when a given use actually associates indices of non-conviction and indices of conviction.

The shuttle process is a consequence of the fact that from a semantic perspective, any difficulty to match prosodic form with interpretation must be interpreted as meaning either: a) that the initial semantic classification is wrong ; b) that discrimination remains suboptimal; c) that the use at stake combines (for a reason which has to be determined) semantic features which are normally mutually exclusive.

As for the present study, case c proved to be the case for 6 out of the 7 "faulty" classifications, all of whom resulted being uses in which the "unconvinced" feature was describing correctly the start of the speakers intervention/use whose ending develop into a finally convinced "oui". Because classification is based on the form of "oui" itself, it may thus be said that the apparent "mistake" was no mistake and instead that it is the initial semantic classification which was suboptimal, illustrating the constant reality that taking into account prosodic form allows for better semantics.

Finally, the 'shuttle' procedure permits displaying two classification rates. The first classification rate is an 'apparent' one and corresponds to the actual rate given by the classification system. However, after careful re-examination and re-interpretation of the errors, the occurences can be re-qualified in cases where the classification machine 'fails rightly' or 'fails wrongly'. In the 'rightly' case, the errors either reflect complex situations which cannot be entirely characterized by prosodic patterns or correspond to situations where co-textual criteria cannot be considered in the prosodic based classifier. Thus, the second classification rate can be introduced as a 'real' machine classification rate which is evaluated after possible relabeling or withdrawing of the misclassified occurences.

4 CONCLUSIONS

In this paper, we have shown that using prosodic information of the word 'oui' can be useful for its semantic interpretation since about 80% of classification rate can be obtained in the classification task between 'conviction' and 'lack of conviction'. Moreover, we have also proved that a feature selection step could improve the classification performance for both the speaker dependent and speaker independent investigated modes. This result can be explained by the curse of dimensionality phenomenon caused by the limited size database.

This study suggested that the semantic interpretation with prosodic analysis could be refined through the 'shuttle' process which consists in reconsidering the misclassified cases possibly 'failing rightly' or 'wrongly'.

Future work concerns the influence of the age and gender of the speaker.

ACKNOWLEDGEMENTS

This work is funded by the région Centre, France. This collaborative work implies four laboratories of the University of Orléans (LLL, PRISME/IRAuS, LIFO, MAPMO). All the persons involved in this project are acknowledged for their active participation.

REFERENCES

Boersma, P. & Weenink, D., 2014. *Praat: doing phonetics by computer.* (Online) Available at: www.praat.org

(Accessed 2014).
Brown, G., Pocock, A., Zhao, M.-J. & Lujan, M., 2012. Conditional Likelihood Maximisation: A Unifying Framework for. *Journal of Machine Learning Research,* Issue 13, pp. 27-66.
Fernandez, R. & W. Picard, R., 2002. *Dialog Act Classification from Prosodic Features Using Support Vector Machines.* Aix-en-Provence, France, s.n., pp. 291-294.
Hasegawa-Johnson, M. & all, &., 2004. *Speech recognition models of the interdependence among syntax, prosody and segmental.* s.l., s.n.
Jain, A., Duin, R. & Mao, J., 2000. Statistical pattern recognition: a review. *Trans. Pattern Analysis and Machine Intelligence,* Jan, 22,(1), pp. 4-37.
Juslin, P. N. & Laukka, P., 2003. Communication of emotions in vocal expression and music performance. *Psychological Bulletin,* 129(5).
Kompe, R., 1997. Prosody in Speech Understanding Systems. *LNAI*, Volume 1307.
Manganaro, L., Peskin, B. & Shriberg, E., 2002. *Using prosodic and lexical information for speaker.* s.l., s.n.
Mary, L. & Yegnanarayana, B., 2008. Extraction and representation of prosodic features for language and and speaker recognition. *Speech Communication,* p. 782–796.
Petit, M., 2009. *Discrimination prosodique et représentation du lexique : application aux emplois des connecteurs discursifs,* s.l.: PhD Thesis, University of Orléans.
Rosenberg, A., 2009. *Automatic Detection and Classification of Prosodic Events,* s.l.: PhD Thesis, University of Columbia.
Shriberg, E. & all, 1998. *Can Prosody Aid the Automatic Classification of Dialog Acts,* s.l.: s.n.
Shriberg, E. & Stolcke, A., 2004. *Direct Modeling of Prosody: An Overview of Applications in Automatic Speech Processing.* s.l., s.n., pp. 575-582.
Spiegl, W. & all, 2009. *Analyzing Features for Automatic Age Estimation on Cross-Sectional Data.* s.l., s.n., pp. 2923-2926.
Szaszák, G., Sztahó, D. & Vicsi, K., 2009. *Automatic Intonation Classification for Speech Training Systems.* s.l., s.n.
Wang, C., 2001. *Prosodic modelling for improved speech recognition and understanding,* s.l.: PhD Thesis, Massachussetts Institute of Technology.
Young, S., Kershaw, D., Odell, J. & Ollason, D., 1999. *The HTK Book.* Cambridge: Entropic Ltd.