Local and Global Feature Selection for Prosodic Classification of the Word's Uses

Abdenour Hacine-Gharbi[1], Philippe Ravier[2] and François Némo[3]
[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
[3]Laboratoire Ligérien de Linguistique, University of Orleans, Orleans, France
{philippe.ravier, francois.nemo}@univ-orleans.fr, gharbi07@yahoo.fr

Keywords: Prosodic Classification, Local and Global Prosodic Features, Dimensionality Reduction, Curse of Dimensionality, Filter Feature Selection, Mutual Information, Classification of Word's Uses.

Abstract: The aim of this study is to evaluate the ability of local or global prosodic features in achieving a classification task of word's uses. The use of French word "oui" in spontaneous discourse can be identified as belonging to the class "convinced (CV)"or "lack of conviction (NCV)". Statistics of classical prosodic patterns are considered for the classification task. Local features are those computed on single phonemes. Global features are computed on the whole word. The results show that 10 features completely explain the two clusters CV and NCV carried out by linguistic experts, the features having being selected thanks to the Max-Relevance Min-Redundancy filter selection strategy. The duration of the phoneme /w/ is found to be highly relevant for all the investigated classification systems. Local features are predominantly more relevant than global ones. The system was validated by building classification systems in a speaker dependent mode and in a speaker independent mode and also by investigating manual phoneme segmentation and automatic phoneme segmentation. In the most favorable case (speaker dependent mode and manual phoneme segmentation), the rate reached 87.72%. The classification rate reached 78.57% in the speaker independent mode with automatic phoneme segmentation which is a system configuration close to an industrial one.

1 INTRODUCTION

Speech communication is a rich but complex way for exchanging information from a speaker to a listener. Indeed, among other modalities, speech brings with itself the information of intonation or prosody. Many studies highlighted the significant role of prosody in syntactic, semantic and pragmatic interpretation (Kompe, 1997) as well as automatic speech recognition (Wang, 2001), speaker identifica-tion (Manganaro et al., 2002), language recognition (Mary and Yegnanarayana, 2008). More recently, since prosody can help refining semantics of the messages, this new information modality has found great interest for semanticists who started to consider that the prosodic features could play a key role in the interpretation and classification of different word's uses (Petit, 2009). Prosodic classification can find a particular interest in oral surveys and opinion polls industry where prosody delivers information that written language cannot. The goal is to exploit this finer information given by key-words in an automatic way for the processing of the large databases at stake.

In a previous paper (Hacine-Gharbi et al., 2015), preliminary results were obtained as an attempt to automatically classify the uses of the French "oui" in a class label "conviction" (CV) or class label "lack of conviction" (NCV). To that aim, a small vocal questionnaire inspired from opinion polls has been created and allowed to collect 118 occurrences for both classes of 'oui'. The classification was performed by a prosodic feature extraction stage combined with a feature selection stage with a "wrapper" strategy. The classifier uses hidden Markov models (HMM) which inputs are the prosodic feature vectors. Each vector is composed of the energy and pitch with the dynamic parameters delta and delta-delta.

The purpose of this work is (i) to enrich the word representation by the individual word duration; (ii) to evaluate how the features are relevant for explaining the two clusters CV and NCV of the full database manually carried out by the linguistic experts; (iii) to investigate the role of local and global prosodic parameters in the classification process. The motivation for the inclusion of this new parameter

Hacine-Gharbi, A., Ravier, P. and Nemo, F.
Local and Global Feature Selection for Prosodic Classification of the Word's Uses.
DOI: 10.5220/0006251407110717
In *Proceedings of the 6th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2017)*, pages 711-717
ISBN: 978-989-758-222-6

comes from the fact that this parameter can be a good discriminator for the two considered classes. This parameter cannot be directly included in a HMM based classifier which requires a sequence of prosodic feature vectors describing local behaviors. This means that the HMM based classifier must be changed which motivates the point (ii). Words have then to be represented by a single vector instead of a sequence of vectors making it possible to introduce the word duration within the single prosodic feature vector. We therefore consider statistics of the prosodic feature vectors. These statistics are the classical mean and standard deviation computed on the feature vectors sequence for each component. However, this can be a sensible strategy only if data are stationary and this is not the case for single words. So the idea is to consider the so-called local features which are computed on the quasi-stationary phonemes which constitute each word. The word “oui” can be decomposed in the phonemes /w/ and /i/. The feature vector is thus the concatenation of the prosodic feature statistics of both phonemes augmented by their respective duration.

Moreover, we are also interested to know whether if the local strategy is relevant. We thus also include the global features, i.e. the prosodic statistics computed on the whole word, in order to evaluate the effect of local vs global prosodic features strategy in the classification system. The answer to this question can be obtained by performing a feature selection procedure which will point out either local or global features within the most relevant ones. Additionally, the proposed study makes the feature selection procedure becoming mandatory because a rather small database size tends to make the rate of classification drop with the increasing vector size. The peaking phenomenon suggests that an optimal number of features have to be selected depending on the size of the database and on the selected features as well. To avoid the well known curse of dimensionality phenomenon (Jain et al., 2000), dimensionality reduction is necessary. The reduction can be achieved thanks to the combination of “greedy” algorithms useful for the feature selection and mutual information measure which is a powerful tool for evaluating the shared information between variables.

The paper is organized as follows: in section 2, the prosodic classification system of the word “oui” is proposed and the construction of the statistical prosodic features vector is detailed. In section 3, the procedure that selects the relevant features for a classification task is presented. Section 4 presents experiments and results in terms of classification rates and selected features using our database. Section 5 concludes the paper.

2 PROSODIC CLASSIFICATION OF THE WORD “OUI”

2.1 Classification System

In the present work, the prosodic pattern is a vector of statistical features (mean, standard deviation) belonging to the class CV or NCV. An appropriate powerful classifier for binary classification is a support vector machine (SVM) classifier. Figure 1 presents the outline of our automatic classification system of prosodic patterns into word’s use.

A SVM based classification system is classically composed of two distinct 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 prosodic features vectors. Then, each sequence is transformed into one vector of statistical features (mean, standard deviation).

First, during the training phase, the system learns occurrences of the training database: the set of statistical features of the training corpus is used for estimating the support vector parameters of the SVM classifier. The “svmtrain” MATLAB command is used with the Kernel function fixed as “rbf” (Gaussian Radial Basis Function kernel). This command uses an optimization algorithm to find support vectors $\mathrm{s_i}$, weights $\mathrm{w_i}$, and bias B that are used to classify unknown prosodic vector x.

Second, during the testing phase, the set of statistical features of the training corpus are the input of the SVM classifier that gives the recognized class using the decision equation (command “svmclassify”):

$$\mathrm{h} = \sum_{\mathrm{i}} \mathrm{w_i} \mathrm{K}(\mathrm{s_i}, \mathrm{x}) + \mathrm{B}$$

where K is the kernel function. If $\mathrm{h} \geq 0$, then x is classified as a CV class, otherwise it is classified as NCV class. The final result is an SVM model that discriminates the two classes through their statistical feature vector. The quality of the classification system is evaluated by a classification rate defined as $\mathrm{TC} = \frac{\mathrm{N-S}}{\mathrm{N}}$, where N is the total number of occurrences given at the input of the classifier and S is the number of misclassified occurrences. The description of the prosodic features is now given in the next section.

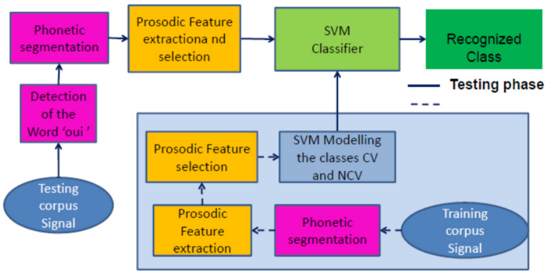
Phonetic segmentation
Prosodic Feature extractiona nd selection
SVM Classifier
Recognized Class
Testing phase
Detection of the Word 'oui '
Testing corpus Signal
Prosodic Feature selection
SVM Modelling the classes CV and NCV
Prosodic Feature extraction
Phonetic segmentation
Training corpus Signal

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

2.2 Prosodic Feature Extraction

Typical features that characterize prosody are the energy Δ (dB) and the pitch F_0 (Hz). Thanks to PRAAT software (Boersma and Weenink, 2014), these parameters are computed every 10 ms on 30 ms analyzing windows of the temporal speech 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 ΔΔ using HTK (Hidden Markov Model Toolkit) library (Young et al., 1999). Thus, each occurrence of the word "oui" is represented by a sequence of vectors with 6 prosodic components noted as $E,\ F_0, \Delta E, \Delta F_0, \Delta\Delta E, \Delta\Delta F_0$. Then the statistical features (mean and standard deviation) of each of the six features are estimated from the sequence. The number of features is therefore 13 when also considering the duration of the sequence. Now, we define the sequences for three segments: the two local phonetic segments of phonemes /w/ and /i/ plus the global segment of the word "oui". The concatenation of the segment vectors produces a statistical prosodic vector composed of 39 features. This vector describes each word occurrence and this is the input of the classification system. The output is the class label CV or NCV which is known during the training phase and induced during the testing phase.

However, the construction of such 39 prosodic statistical features vector basically requires the detection of the words "oui" as well as the segmentation of each detected word in the two phonemes /w/ and /i/. These tasks have been manually achieved with the help of two software: PRAAT for the detection of words from spontaneous speech discourses and Easyalign for the phonetic segmentation. In this work, 115 words have been segmented by Easyalign and manually corrected. These steps of detection and segmentation can be achieved using HMM segmentation but we have preferred a manual segmentation for guarantying the best segmentation. Indeed, the best accuracy of some measures like duration of the word or phonemes is also needed as new local and global features in this study. However, in order to be as near as possible from a real classification system, we investigated an automatic segmentation system based on the isolated words we have constructed. The performance of both classifications systems will be discussed in section IV. Note also from fig. 1 that the system mentions a "prosodic feature selection" stage for the training phase whereas and a "selected prosodic features" stage is mentioned for the testing phase. This means that training is useful for selecting relevant prosodic features used for the testing phase. The feature selection procedure may either belong to the "wrapper" strategy or to the "filter" strategy. With the "wrapper" strategy, the feature selection is performed upon a classification performance rate criterion. This necessitates designing a new classifier each time a feature vector structure is tested. This is a prohibitive task when the database is growing or when the number of candidate vectors subsets to be selected is very large. Since the application of this work concerns 39 features giving rise to the peaking phenomenon and necessitating the feature selection), we therefore prefer the "filter" strategy that permits to select the features that best describe the classes. Indeed, the combinatorial possibilities of candidate subsets are very huge. The "filter" strategy is only based on the information given by the feature vector of each occurrence with its belonging class without having to consider any classification method as required for "wrapper" strategy. The feature selection stage is described in the next section.

3 PROSODIC FEATURE SELECTION

Feature selection consists in choosing a subset S_{opt} of k features $\{Y_{P_1}, Y_{P_2} \cdots, Y_{P_k}\}$ from a set F of n features $\{Y_1, Y_2 \cdots, Y_n\}$ such that S_{opt} keeps most of the information useful for a classification task. The quantity of information brought by the subset S_{opt} is often evaluated thanks to the mutual information (MI) measure because of its ability of assessing the nonlinear statistical dependency between variables (Cover and Thomas, 1991). So the subset S_{opt} that best describes the classes C (in our cases the class indexes labeled CV or NCV) is the subset that maximizes the MI between S_{opt} and C:

$$S_{opt} = \arg\max_{S \subset F} I(C; S). \tag{1}$$

To circumvent the prohibitive search which becomes intractable when the size of S grows, “greedy forward” search strategies can be employed. The search is an iterative algorithm that proposes at each iteration j the best feature $\mathrm{Y_{P_j}}$ from the unselected features set. This new selected feature is then appended to the already selected subset $\mathrm{S_{j-1}}$ generating $\mathrm{S_j = Y_{P_j} \cup S_{j-1}}$ (Brown et al., 2012):

$$\mathrm{Y_{P_j}} = \arg\max_{\mathrm{Y_i \in F - S_{j-1}}} \left[\mathrm{I}\left(\mathrm{C; Y_i \backslash S_{j-1}}\right)\right] \qquad (2)$$

Equation (2) can also be expanded in a multivariate MI of order 3 between $\mathrm{C, Y_i}$ and $\mathrm{S_{j-1}}$ as:

$$\mathrm{Y_{P_j}} = \arg\max_{\mathrm{Y_i \in F - S_{j-1}}} \left[\mathrm{I(C; Y_i) - I_3\left(C; Y_i; S_{j-1}\right)}\right] \quad (3)$$

The evaluation of $\mathrm{I_3\left(C; Y_i; S_{j-1}\right)}$ becomes very difficult when j grows because this evaluation requires the estimation of high-dimensional probability density functions that cannot be precise enough for fixed database sizes. Most of the algorithms propose a simplification of (3) following different strategies like MIM, MIFS, MRMR, CMI, DISR, CIFE, TMI, ICAP (Brown et al., 2012). In (Brown et al., 2012), the authors conclude that the JMI strategy provides a good compromise between precision, flexibility and stability when the database is small size. They also point out the MRMR and CMI strategies that better perform than other ones in terms of balance between high relevance and small redundancy. We give below the derivation of (3) for the three selected strategies (Brown et al., 2012).

- MRMR (Max-Relevance Min-Redundancy)

$$\mathrm{Y_{P_j}} = \arg\max_{\mathrm{Y_i \in F - S_{j-1}}} \left[\mathrm{I(C; Y_i)} - \frac{1}{\mathrm{j-1}} \sum_{\mathrm{k=1}}^{\mathrm{j-1}} \mathrm{I\left(Y_i; Y_{p_k}\right)}\right] \quad (4)$$

- JMI (Joint Mutual Information)

$$\mathrm{Y_{P_j}} = \arg\max_{\mathrm{Y_i \in F - S_{j-1}}} \left[\mathrm{I(C; Y_i)} - \frac{1}{\mathrm{j-1}} \sum_{\mathrm{k=1}}^{\mathrm{j-1}} \mathrm{I_3\left(C; Y_i; Y_{P_k}\right)}\right] (5)$$

- CMI (Conditional Mutual Information)

$$\mathrm{Y_{P_j}} = \arg\max_{\mathrm{Y_i \in F - S_{j-1}}} \left[\mathrm{I(C; Y_i)} - \max_{\mathrm{Y_{P_k} \in S_{j-1}}} \mathrm{I_3\left(C; Y_i; Y_{P_k}\right)}\right] \quad (6)$$

For JMI and CMI strategies, the term $\mathrm{I_3\left(C; Y_i; Y_{P_k}\right)}$is actually computed as $\mathrm{I\left(Y_i; Y_{P_k}\right) - I\left(Y_i; Y_{P_k} \backslash C\right)}$. The MI $\mathrm{I(X; Y)}$ between variables X and Y is expressed as $\mathrm{I(X; Y)} = \iint_{-\infty}^{+\infty} p(x,y) log\left(\frac{p(x,y)}{p(x)p(y)}\right) dxdy$ where $p(x,y)$is the joint distribution of (X,Y) and $p(x)$ and $p(y)$are the marginal distributions . This continuous definition can be estimated by discretization of the $\mathrm{I(X; Y)}$ formula applying the histogram low mean square error estimation as described in (Hacine-Gharbi et al., 2013).
The three strategies have been used and compared for the feature selection procedure.

4 EXPERIMENTS AND RESULTS

Three experiments are conducted in order (i) to evaluate the relevance of the type of features, either local or global; (ii) to know whether a set of selected features is able to completely explain the two clusters composed of 64 CV words and 51 NCV words labeled by the linguistic experts; (iii) to evaluate the classification rates in real situations where feature selection and training phase are performed on the split database. In this last case, the phonetic segmentation has been achieved either manually by visual inspection or automatically by modeling the word ‘oui’ into the phonemes /w/ and /i/. Each phoneme is modeled by a 3 states HMM. Each state is modeled by a Gaussian mixture model (GMM) with three Gaussians. The implementation of the system is carried out using the HTK library (Young et al., 1999). This task helps identifying the begin and the end of each phoneme which permits to calculate the local parameters in the phoneme regions. The performance of this segmentation is compared to the performance obtained with a manual segmentation in terms of precision or classification rate.

4.1 Performance Study using Local and Global Features

This experiment permits to analyze the pertinence of local or global prosodic features that should be able to explain the index variable representing the class each occurrence belongs to. The same database of 115 occurrences was considered both for the training phase and the testing phase. The goal in this section is to evaluate whether the features are pertinent for explaining the two clusters of CV and NCV words proposed by the linguistic experts. The performance study is achieved by independently considering three sets of features: local features are referred to the statistical prosodic features for the phoneme /w/

Table 1: Performance comparison of global (GF) and local features (LFi and LFw).

	μ_E	μ_{F_0}	$\mu_{\Delta E}$	$\mu_{\Delta F_0}$	$\mu_{\Delta\Delta E}$	$\mu_{\Delta\Delta F_0}$	σ_E	σ_{F_0}	$\sigma_{\Delta E}$	$\sigma_{\Delta F_0}$	$\sigma_{\Delta\Delta E}$	$\sigma_{\Delta\Delta F_0}$	d
LFw	70.44	78.26	79.13	87.82	93.04	92.17	93.91	93.91	96.52	97.39	98.26	98.26	98.26
LFi	60.87	71.30	74.78	81.74	85.22	84.35	86.09	87.83	91.30	87.83	87.83	86.09	89.57
GF	58.26	73.91	73.91	80.87	86.08	88.70	93.04	94.78	95.65	95.65	96.52	95.65	96.52

Table 2: Classification rates obtained with feature selection procedure and the selected feature names at each iteration.

	1	2	3	4	5	6	7	8	9	10	11	12	13
MRMR	76.52 $d^{(w)}$	80.87 $\mu_{\Delta F_0}^{(i)}$	88.70 $\mu_E^{(w)}$	88.70 $\sigma_{\Delta\Delta E}^{(i)}$	93.04 $\mu_{F_0}^{(i)}$	94.78 $\mu_{\Delta\Delta F_0}^{(w)}$	95.65 $\mu_{\Delta\Delta E}^{(w)}$	98.26 d^G	99.13 $\sigma_{F_0}^{(w)}$	100 $\sigma_{\Delta\Delta E}^{(w)}$	100 $\mu_E^{(i)}$	100 $\mu_{\Delta\Delta E}^{(i)}$	100 $\mu_{\Delta F_0}^{(w)}$
JMI	76.52 $d^{(w)}$	80.87 $\mu_{\Delta F_0}^{(i)}$	82.61 d^G	84.35 $d^{(i)}$	80.87 $\mu_{\Delta\Delta F_0}^{(w)}$	93.91 $\mu_E^{(w)}$	93.04 $\mu_{\Delta\Delta E}^{(i)}$	93.91 $\mu_{\Delta E}^{(G)}$	93.91 $\mu_{\Delta E}^{(i)}$	96.52 $\sigma_{\Delta\Delta F_0}^{(w)}$	96.52 $\mu_{\Delta F_0}^{(w)}$	96.52 $\sigma_{F_0}^{(w)}$	96.52 $\sigma_{F_0}^{(w)}$
CMI	76.52 $d^{(w)}$	80.87 $\mu_{\Delta F_0}^{(i)}$	88.70 $\mu_E^{(w)}$	86.96 d^G	87.83 $\sigma_{\Delta\Delta E}^{(i)}$	88.70 $\mu_{\Delta E}^{(i)}$	93.04 $\sigma_{F_0}^{(i)}$	93.04 $\sigma_{\Delta\Delta F_0}^{(G)}$	96.52 $\mu_{\Delta F_0}^{(w)}$	96.52 $\sigma_{\Delta\Delta F_0}^{(w)}$	97.39 $\sigma_{\Delta F_0}^{(G)}$	98.26 $\mu_E^{(i)}$	98.26 $\mu_{\Delta E}^{(G)}$

(LFw) or the phoneme /i/ (LFi) and global features are referred to the whole word “oui”. The results expressed in terms of the classification rate are shown in Table I, in the natural order of the 13 parameters composed of 6 mean values $\mu_E, \mu_{F_0}, \mu_{\Delta E}, \mu_{\Delta F_0}, \mu_{\Delta\Delta E}, \mu_{\Delta\Delta F_0}$ and 6 standard deviation values $\sigma_E, \sigma_{F_0}, \sigma_{\Delta E}, \sigma_{\Delta F_0}, \sigma_{\Delta\Delta E}, \sigma_{\Delta\Delta F_0}$ plus the duration d of the segment. This table highlights the pertinence of the local features corresponding to the first phoneme /w/ since values are always higher for LFw. The LFi values are less informative for the classification task since the classification rates are in average 8% below those obtained for LFw. However, the 13 available features do not explain 100% of the class indexes. It is thus necessary to increase the number of features by concatenating them.

4.2 Feature Selection Results

This part is devoted to the selection of the most relevant features explaining the classes or clusters CV and NCV among the 39 local and global features. The complete database with the 115 occurrences is employed for the selection. The classification rates are evaluated on the same database. Results are reported in table II for the 13 first selected features. Each reported classification rate mentions the new selected feature by one of the three tested strategies (MRMR, JMI or CMI). The features are mean $\boldsymbol{\mu}$, standard deviation $\boldsymbol{\sigma}$ or duration $\mathbf{d}$. Superscript conventions are adopted for indicating a local or global feature: the notations $(\boldsymbol{w})$ and $(\boldsymbol{i})$ are referred to a local feature for phonemes /w/ and /i/ respectively; the notation $\boldsymbol{G}$ is referred to a global feature.

The results in table II show that the 10 first features selected from 39 by the MRMR strategy can explain both clusters manually carried out by the linguistic experts since the classification rate is 100%. Moreover, the MRMR strategy always gives the best results comparatively to JMI and CMI. The most interesting result is the selection of the duration of the phoneme /w/ as first feature, whatever the strategy. In addition, the dynamic pitch feature of the second phoneme is a major feature since always selected as second one. The pitch is relevant when observed through its temporal domain variations expressed by $\boldsymbol{\mu}_{\Delta F_0}^{(i)}$. The results also show that the global duration parameter is always selected among the 10 first parameters whatever the strategy. Note an other important result suggesting that local features are more relevant than global ones since the 10 first selected features are predominantly local.

We analyzed in table III the role of the segment duration features (DF). Classification rates are shown with single feature and the combinations of two features from 3, plus the 3 features. The role of $\mathbf{d}^{(w)}$ is confirmed as first feature. Moreover, the combination of two duration features also requires $\mathbf{d}^{(w)}$ to be selected. The conclusion is that this parameter is highly relevant.

Table 3: Classification rates obtained with duration features.

	$\{d^{(w)}\}$	$\{d^{(i)}\}$	$\{d^G\}$	$\{d^{(w)}, d^{(i)}\}$ or $\{d^{(w)}, d^G\}$	$\{d^{(i)}, d^G\}$	$\{\mathbf{d}^{(w)}, \mathbf{d}^{(i)}, \mathbf{d}^{G}\}$
DF	76.52	64.35	70.43	85.22	80.86	85.22

4.3 Classification Rates in a Real Context

The classification has been evaluated on dependent speaker mode and on independent speaker mode. In the first mode, the experiment evaluates the classification rates when classically splitting the database in a training part (58 occurrences, 27 CV and 31 NCV) and a testing part (57 occurrences, 37 CV and 20 NCV). The occurrences of each speaker were dispatched into both parts (speaker dependent system). Two classification systems were built: the first one investigated a manual phoneme segmentation achieved by linguistic experts while the second one investigated automatic phoneme segmentation. For each system, the classification rates are displayed in Fig. 2 and Fig. 3 as a function of the number of selected features, using one strategy among MRMR, JMI or CMI.

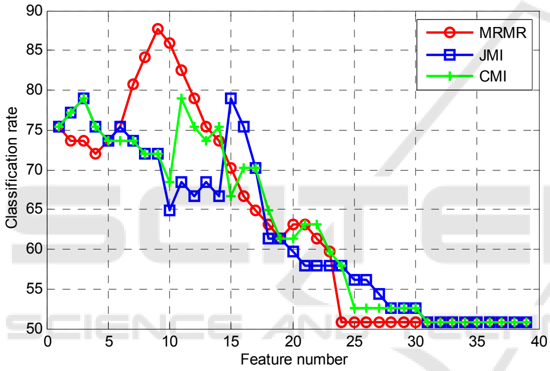

Figure 2: Classification rates with feature selection achieved on the 58 occurrences with manual segmentation in speaker dependent mode.

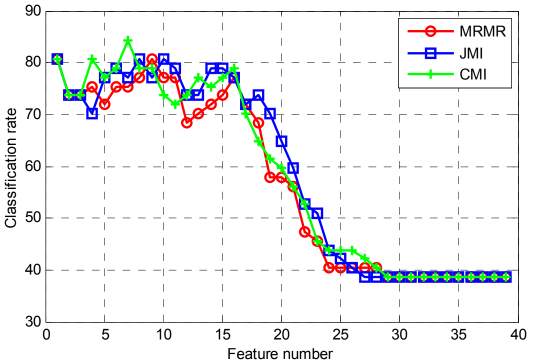

Figure 3: Classification rates with feature selection achieved on the 58 occurrences with automatic segmentation in speaker dependent mode.

Both graphs show that the peaking phenomenon which is due to the relatively too small number of occurrences for the training phase (58) regarding the potential total number of features (39). The best results give a classification rate of 87.72% with 9 features and MRMR strategy in the manual segmentation case and 84.21% in the automatic segmentation case with 7 features and CMI strategy, respectively. The three selection methods gave the same first selected feature (duration of the /w/ phoneme), whatever the segmentation procedure. In both cases, the feature selection process permits to limit the number of features to be considered for classification.

In the second mode, the experiment evaluates the classification rates when classically splitting the database in a training part (59 occurrences, 33 CV and 26 NCV) and a testing part (56 occurrences, 31 CV and 25 NCV).

In this mode, the speakers participating in the testing phase have not participated in the training phase. Two classification systems were similarly built to the independent case. For each system, the classification rates are displayed in Fig. 4 and Fig. 5 as a function of the number of selected features, using one strategy among MRMR, JMI or CMI.

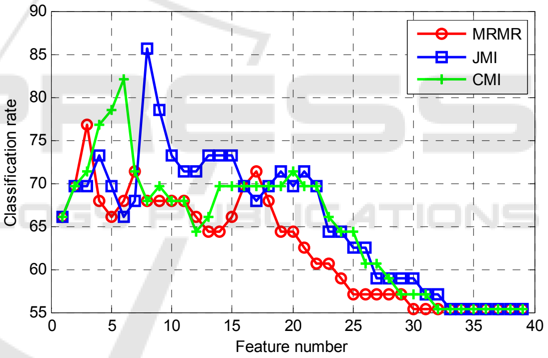

Figure 4: Classification rates with feature selection achieved on the 59 occurrences with manual segmentation in speaker independent mode.

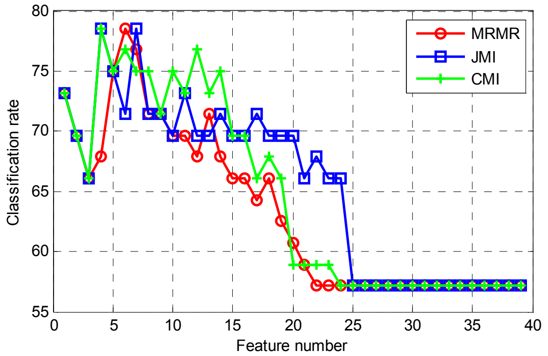

Figure 5: Classification rates with feature selection achieved on the 59 occurrences with automatic segmentation in speaker independent mode.

Like in the dependent mode, the three selection methods highlight the /w/ phoneme duration as a relevant feature since all the methods firstly select this feature. The maximum rate for the manual segmentation (85.71% with JMI strategy and 8 features) is near from the rate previously achieved in the dependant mode. As expected, a classification rate decrease is observed in the independent mode with automatic segmentation which however remains acceptable when considering this system close from an industrial system (maximum of 78.57% whatever the feature selection method).

Nevertheless, these results remain partial ones since a real functional system requires a large database which was not available in our study.

5 CONCLUSION

The aim of this study was to evaluate the ability of local or global prosodic features for explaining the two clusters carried out by linguistic experts and for classifying the word's uses of "oui" in a real context of spontaneous discourse. The word's uses were identified as belonging to the "convinced" or "lack of conviction" class. The results showed that 10 features are sufficient to fully explain both clusters CV and NCV based on the 115 occurrences of a self made corpus which have been labeled by linguistic experts; the features having being selected thanks to the MRMR filter selection strategy. The 10 relevant selected features by this strategy are local for the most part. All the results showed that the first relevant feature was the /w/ phoneme duration. The system was validated by building classification systems in a speaker dependent mode and in a speaker independent mode and by also investigating manual phoneme segmentation and automatic phoneme segmentation. In the case of speaker dependent mode and manual phoneme segmentation, the rate reached 87.72%. The classification rate reached 78.57% in the speaker independent mode with automatic phoneme segmentation which is a system configuration close to an industrial one. These results are partial and preliminary ones regarding the size of the database. However, these are promising for industrial applications like automatic processing of large database oral opinion polls.

ACKNOWLEDGMENT

This work is funded by the région Centre Val de Loire, 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. and Weenink, D., 2014. *Praat: doing phonetics by computer*. [Online] (5.3.75) Available at: www.praat.org [Accessed 2014].

Brown, G., Pocock, A., Zhao, M.-J. and Lujan, M., 2012. Conditional Likelihood Maximisation: A Unifying Framework for information theoretic feature selection. *Journal of Machine Learning Research*, 13(1), pp.27-66.

Cover, T. and Thomas, J., 1991. *Elements of information theory*. New York: Wiley Series in telecommunications.

Hacine-Gharbi, A. et al., 2013. A new histogram-based estimation technique of entropy and mutual information using mean squared error minimization. *Computers and Electrical Engineering*, 39(3), pp.918-33.

Hacine-Gharbi, A., Petit, M., Ravier, P. and Nemo, F., 2015. Prosody Based Automatic Classification of the Uses of French 'oui' as Convinced or Unconvinced Uses. In *4th International Conference on Pattern Recognition Applications and Methods (ICPRAM)*. Lisboa, Portugal, 2015.

Jain, A., Duin, R. and Mao, J., 2000. Statistical pattern recognition: a review. *Trans. Pattern Analysis and Machine Intelligence*, 22, (1), pp.4-37.

Kompe, R., 1997. Prosody in Speech Understanding Systems. *LNAI*.

Manganaro, L., Peskin, B. and Shriberg, E., 2002. Using prosodic and lexical information for speaker. In *ICASSP*., 2002.

Mary, L. and Yegnanarayana, B., 2008. Extraction and representation of prosodic features for language and speaker recognition. *Speech Communication*, pp.782–96.

Petit, M., 2009. *Discrimination prosodique et représentation du lexique: application aux emplois des connecteurs discursifs*. thesis. PhD Thesis, University of Orléans.

Wang, C., 2001. *Prosodic modelling for improved speech recognition and understanding*. PhD Thesis, Massachussetts Institute of Technology.

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