Importance of Prosody for Dialogue Acts Recognition

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

[1]LORIA UMR 7503, BP 239 - 54506 Vandoeuvre, FRANCE

[2]Dept. Informatics & Computer Science, University of West Bohemia, Plzeň, Czech Republic

{kral,cerisara}@loria.fr,kleckova@kiv.czu.cz

Abstract

This paper deals with automatic Dialogue Acts (DAs) recognition in French and in Czech. In this work, only prosodic features are considered. The utterances are recognized according to three types of dialogue acts: statements, yes/no questions and other questions, mainly wh-questions. We show that it is not possible to recognize all utterances only with basic prosodic features (F0 and energy) in real conditions with a good accuracy.

1. Introduction

This work aims at dialogue acts recognition. We investigate the possibility to recognize dialogue acts (statements and questions) in French and in Czech from prosodic features only. The objective of this work is to study the possible use of prosodic features to automatically recognize three classes of DAs: statements, yes/no questions and other questions (wh-question, etc.) in French and in Czech. This DA tag-set is derived from the seven classes considered in Section 2, which have been further simplified with regard to the specifics to our needs and to contains of our corpora.

Different types of information (lexical, syntax, dialogue grammar, etc.) can be used to recognize dialogue acts. In this work, we consider only two prosodic features: the Fundamental Frequency (F0) and the energy. The next versions of the system will integrate other knowledge sources.

After introducing to our research domain, the existing studies in order to recognize the DAs are shown. Then, the information about our DAs corpora is given. In the following section, the DAs for French and for Czech are recognized and both corpora are analyzed. The results of experiments are concluded in the last part of this paper.

2. Short Review of Dialogue Acts Recognition Approaches

To the best of our knowledge, there are very few existing work on automatic modeling and recognition of dialogue acts in the French and Czech language. Alternatively, a number of studies have been published for other languages, and particularly for English and German.

In most of these works, the first step is to define the set of dialogue acts to recognize. In [1, 2, 3], 42 dialogue acts classes are defined, based on the Discourse Annotation and Markup System of Labeling (DAMSL) tag-set [4]. This list is usually reduced into a much smaller number of broad classes, because some classes occur only seldom, and because all these classes are not needed for dialogue understanding. A common regrouping is the following [1]:

- statements
- questions
- backchannels
- incomplete utterance

- agreements
- appreciations
- other

Automatic recognition of these dialogue acts can then be achieved using one of, or a combination of the three following models:

- DA-specific language models
- DA-specific prosodic models
- dialogue grammar

The first class of models infers the DA associated to an utterance from its words sequence. It generally uses probabilistic language models such as n-gram [2, 5], semantic classification trees [5], or neural networks [6, 7]. This lexical information usually contributes the most to characterize the utterance DA.

Prosodic models are often used to provide additional clues to classify utterances in terms of DAs. The dialogue acts can be characterized by prosody as follows [8]:

- a falling intonation for a statement
- a rising F0 contour for a question (particularly for declaratives and yes/no questions)
- a continuation-rising F0 contour characterizes a (prosodic) clause boundaries, which is different from the end of utterance

Prosody is used for DAs recognition in [1, 2, 3, 9, 10, 11]. In [1], the duration, pause, fundamental frequency, energy and speaking rate are modeled by a CART-style decision trees classifier. In [9], prosody is used to segment utterance. The duration, pause, F0-contour and energy features are used in [10, 11]. These two studies compute several features based on these basic prosodic attributes, for example the max, min, mean and standard deviation of F0, the mean and standard deviation of the energy, the number of frames in utterance and the number of voiced frames. The features are computed on the whole utterance and also on the last 200 ms of each utterance. The authors conclude that the end of utterances carry the most important prosodic information for DAs recognition. Furthermore, three different classifiers: hidden Markov models, classification and regression trees and neural networks are compared, and give similar DAs recognition accuracy.

Very often, a dialogue grammar is further used to predict the most probable next dialogue act based on the previous ones. It can be modeled by hidden Markov models [2, 3] or Discriminative Dynamic Bayesian Networks (DBNs) [12].

The lexical and prosodic classifiers are combined in [1, 2, 3]. The following equation is used:

$$P(W,F|C) = P(W|C).P(F|W,C) \approx P(W|C).P(F|C) \quad (1)$$

where C represents a dialogue act and W and F represents respectively the lexical and prosodic information. W and F are assumed independent.

3. Dialogue Acts Corpora

For French, we work on the ESTER corpus [13], which contains natural human-human speech from French broadcast news evaluation. This corpus has not been designed *a priori* to do DAs recognition DAs have been manually annotated. Our French DAs corpus contains 746 utterances: 259 statements (S), 231 yes/no questions (Q[y/n]) and 256 other questions (Q).

For Czech, the Czech Railways corpus is used. It contains human-human dialogues in ticketing reservation task. This corpus was created at the University of West Bohemia mainly by members of the Department of Computer Science and Engineering. For our experiments, one subset of this corpus was labelled manually according to three DAs. Our Czech DAs corpus contains 862 utterances: 290 statements (S), 282 yes/no questions (Q[y/n]) and 290 others questions (Q).

The class *questions* has been separated in both languages into two classes: yes/no questions and other questions. The yes/no question is a question which has usually an answer: “yes” or “no”. It is often characterized by increasing F0 contour [14]. Other questions (mostly wh-questions) are usually identified by an interrogative word (wh-word) and their melody is increasing not any time. It will be probably difficult to recognize this type of DAs by prosody only.

All the following experiments are realized using a cross-validation procedure, where 10% of the corpus is reserved for the test, and another 10% for the development set. The resulting global accuracy has a confidence interval < +/- 2.5%.

4. Experiments

4.1. Dialogue acts recognition by prosody

Following the conclusions of previous studies [15, 16], only the two most important prosodic attributes: F0 and energy, and only final segments of the DAs are used. The F0 curve is computed with the autocorrelation function. The F0 and energy values are computed on every overlapping speech window. The F0 curve on the unvoiced parts of the signal is completed by linear interpolation. Then, each utterance is decomposed into 20 segments and the average values of F0 and energy are computed for each segment. This number is chosen experimentally [16]. We thus obtain 20 values of F0 and 20 values of energy per utterance. Let us call F the set of prosodic features for one utterance and C is the dialogue act class. We used for DA recognition a Gaussian Mixture Model (GMM) classifier that models $P(F|C)$.

The best recognition accuracy, which is shown for French in Table 1, is obtained with 3-mixtures GMM. The global accuracy is 51%. The best recognition accuracy for Czech is obtained with 5-mixtures GMM (c.f. Table 2). The global ACC of this experiment is 49%.

The experiments with more Gaussians have a lower accuracy, because of the lack of training data.

Table 1. GMM’s confusion matrix for French language in %

		Recognized class in [%]	
Pronounced class	S	Q	Q[y/n]
S	**45**	36	19
Q	34	**43**	23
Q[y/n]	17	19	**64**

Table 2. GMM's confusion matrix for Czech language in %

		Recognized class in [%]	
Pronounced class	S	Q	Q[y/n]
S	**56**	25	19
Q	30	**37**	33
Q[y/n]	19	27	**54**

4.2. Corpora analysis

We first analyze the F0 curve for all DAs. We compute the mean and variance values for all features. The means values of F0 for French are shown in Figure 1 and for Czech in Figure 2. The variances are not shown, because there are very small (in interval (0; 0,02]), and because the figure would be difficult to be read.

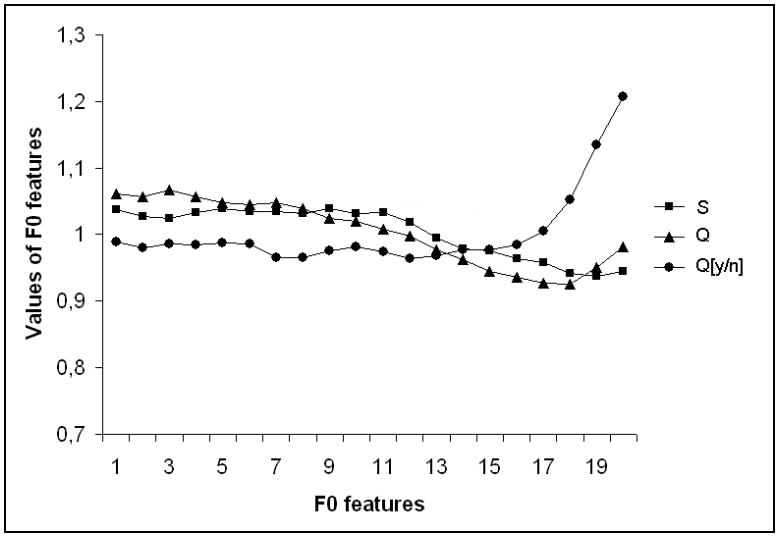

Fig. 1. F0 curves for three types of DAs for French: S=statements, Q=other questions, Q[y/n]=yes/no question

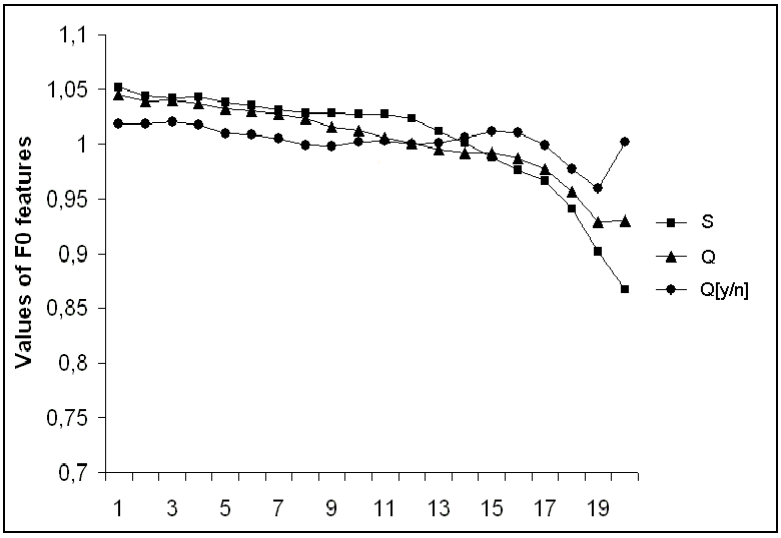

Fig. 2. F0 curves for three types of DAs for Czech: S=statements, Q=other questions, Q[y/n]=yes/no questions

In the first part of the segment the F0 curve is very similar for all DAs. The last third of the segment is the most discriminating for both languages.

For French, the ending F0 slope of yes/no questions (Q[y/n]) is clearly increasing. These can explain the good recognition accuracy for this DA. Conversely, F0 curve for statements (S) and for other questions (Q) is falling or almost horizontal and their values are very close, which leads to some confusion in automatic recognition.

For Czech, the ending F0 slope of yes/no questions (Q[y/n]) is increasing, decreasing for statements (S) and almost horizontal for other questions (Q).

Next, we analyze the F0 slope at the end of the utterance. The objective is to assess the basic prosodic rules described in Section 2. Four values of F0 are computed on the last part on the utterance by an autocorrelation function. Linear regression is performed on these four values.

Table 3 shows the number of utterances according to prosodic rules for French, Table 4 shows the results for the Czech language. The column with the "/" symbol represents the utterances with positive F0 slope and these with the "\" symbol with negative one. This first analysis separates the linear regression values into two intervals only.

Table 3. Analysis of the slope of F0 curve at the end of utterances by linear regression for French DA corpus

Class	\	/	< -0.03	[-0.03; 0,03]	0.03 <
S	60	40	34	42	24
Q	53	47	25	39	36
Q[y/n]	23	77	14	17	69

Table 4. Analysis of the slope of F0 curve at the end of utterances by linear regression for Czech DA corpus

Class	\	/	< -0.03	[-0.03; 0,03]	0.03 <
S	88	12	57	39	4
Q	80	20	42	47	11
Q[y/n]	57	43	32	43	25

In the next analysis, we from thus split the range of the final slope into three segments. In the first one, there are all values of linear regression greater as 0.03. It may be a characteristic for the questions. The second interval is [-0.03; 0.03]. The utterances with linear regression coefficients smaller that -0.03 are in the last column of the table. It may be a characteristic of statements.

5. Conclusion

We show that it is not possible to recognize all utterances only with basic prosodic features (F0 and energy) in real conditions with a good accuracy. It is due to an important overlapping between the features values in the classes.

For French, the most discriminating are yes/no questions, where the accuracy is 64%. The recognition accuracy of other DAs (S and Q) is about 50%. For Czech, the less discriminating are other questions (Q), where the accuracy is 37% only. The recognition accuracy of the other DAs (S and Q[y/n]) is about 55%.

The perspectives will consist to use other models such as DA-specific language models and dialogue grammar, as reviewed in Section 2, to improve the DA recognition accuracy.

Acknowledgements

This work has been partly supported by the European integrated project Amigo (IST-004182), a project partly funded by the European Commission, and by the Ministry of Education, Youth and Sports of Czech republic grant (NPV II-2C06009).

References

1. *E. Shriberg et al.*, "Can prosody aid the automatic classification of dialog acts in conversational speech?," in Language and Speech, 1998, vol. 41, pp. 439–487.

2. *A. Stolcke et al.*, "Dialog act modeling for conversational speech," in AAAI Spring Symp. on Appl. Machine Learning to Discourse Processing, 1998, pp. 98–105.

3. *A. Stolcke et al.*, "Dialog act modeling for automatic tagging and recognition of conversational speech," in Computational Linguistics, 2000, vol. 26, pp. 339–373.

4. *J. Allen and M. Core*, "Draft of DAMSL: Dialog act markup in several layers," 1997.

5. *M. Mast et al.*, "Automatic classification of dialog acts with semantic classification trees and polygrams," in Connectionist, Statistical and Symbolic Approaches to Learning for Natural Language Processing, 1996, pp. 217–229.

6. *H.-F. Wang, W. Gao, and S. Li*, "Dialog acts analysis of spoken chinese based on neural networks," Chinese Journal of Computers, 1999.

7. *T. Andernach, M. Poel, and E. Salomons*, "Finding classes of dialogue utterances with Kohonen networks," in ECML/MLnet Workshop on Empirical Learning of Natural Language Processing Tasks, Prague, Czech Republic, April 1997, pp. 85–94.

8. *R. Kompe*, Prosody in Speech Understanding Systems, Springer-Verlag, 1997.

9. *M. Mast, R. Kompe, S. Harbeck, A. Kiessling, H. Niemann, E. N¨oth, E. G. Schukat-Talamazzini, and V. Warnke*, "Dialog act classification with the help of prosody," in ICSLP'96, Philadelphia, 1996.

10. *H. Wright*, "Automatic Utterance Type Detection Using Suprasegmental Features," in ICSLP'98, Sydney, 1998, p. 1403.

11. *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.

12. *G. Ji and J. Bilmes*, "Dialog act tagging using graphical models," in ICASSP'05, Philadelphia, March 2005.

13. *Département Technologies de l'Information et de la Communication Action Technolangue*, "French ESTER Corpus," in *http://www.recherche.gouv.fr/technolangue/*.

14. *Palková Z.*, Fonetika a fonologie češtiny, Karolinum publisher of Charles University in Prague, 1997.

15. *V. Strom*, "Detection of Accents, Phrase Boundaries and Sentence Modality in German with Prosodic Features," in Eurospeech'95, Madrid, Spain, 1995.

16. *J. Kleckova and V. Matousek*, "Using Prosodic Characteristics in Czech Dialog System," in Interact'97, 1997.