Combining lexical and prosodic features for automatic detection of sentence modality in French

Luiza Orosanu, Denis Jouvet

Combining lexical and prosodic features for automatic detection of sentence modality in French

Luiza Orosanu[1,2,3] and Denis Jouvet[1,2,3]

Speech Group, LORIA
[1] Inria, Villers-lès-Nancy, F-54600, France
[2] Université de Lorraine, LORIA, UMR 7503, Villers-lès-Nancy, F-54600, France
[3] CNRS, LORIA, UMR 7503, Villers-lès-Nancy, F-54600, France
{luiza.orosanu,denis.jouvet}@loria.fr

Abstract. This article analyzes the automatic detection of sentence modality in French using both prosodic and linguistic information. The goal is to later use such an approach as a support for helping communication with deaf people. Two sentence modalities are evaluated: questions and statements. As linguistic features, we considered the presence of discriminative interrogative patterns and two log-likelihood ratios of the sentence being a question rather than a statement: one based on words and the other one based on part-of-speech tags. The prosodic features are based on duration, energy and pitch features estimated over the last prosodic group of the sentence. The evaluations consider using linguistic features stemming from manual transcriptions or from an automatic speech transcription system. The behavior of various sets of features are analyzed and compared. The combination of linguistic and prosodic features gives a slight improvement on automatic transcriptions, where the correct classification performance reaches 72%.

Keywords: speech-to-text transcriptions, question detection, prosody, likelihood ratio, part-of-speech tags

1 Introduction

The automatic detection of sentence modality has been studied in the past decades with different objectives: to model and detect the speech structure [1], to distinguish questions from statements [2–7], to create the summary of documents or meetings [4], to enrich an automatic transcription with punctuation marks [8], etc.

The most useful cues for the detection of sentence modality are the prosodic features (computed over the speech signal) and the linguistic features (computed over the word transcription). There are two scenarios for the linguistic features: when they are extracted from correct data (textual and/or manual transcriptions of audio) or from automatic transcriptions (generated by a speech recognition system). The studies related to automatic speech recognition systems have

to additionally take into account the speech recognition errors which get more frequent for poor sound qualities and on spontaneous speech, and can highly decrease the classification performance.

Regarding the prosodic features, different studies on different languages consider different features computed over different parts of the speech signal. In [2], the prosodic features (pitch and energy) computed on the last 700 milliseconds of speech were used for the detection of French questions. In [6], the energy and the fundamental frequency were the key features in the detection of Arabic questions. In [9] the English question asking behavior was designed in order to improve the intelligent tutoring systems; their study concluded that the most useful features were the pitch slope of the last 200 milliseconds of a turn. Another detector of French questions (versus statements) made use of 12 prosodic features derived from the fundamental frequency of the entire utterance [4].

When dealing with correct data (e.g. manual transcriptions), considering both prosodic and lexical features proves very useful. In [5], the combined prosodic-lexical classifier considers lexical features relative to interrogative terms: the unigrams/bigrams preceding or succeeding interrogative terms and the presence (or absence) of interrogative terms. The use of web textual conversations to detect questions in conversational speech was analyzed in [7]. Their lexical features consider the presence or absence of unigrams through trigrams in the sentences (with respect to questions or statements).

When dealing with automatic transcriptions, the sentence modality detection becomes more challenging. In [1], 42 dialog acts were used to model and detect the discourse structure of natural English speech (human-to-human telephone conversations). They used three different types of information (linguistic, prosodic and a statistical discourse grammar) and achieved an accuracy of 65% on ASR transcripts versus 72% on reference manual transcripts. Combining recognized words with the discourse grammar was the most useful for this task. The detection of questions in English meetings was addressed in [10], using lexico-syntactic, turn related and pitch related information. They achieved an accuracy of 54% on ASR transcripts versus 70% on reference manual transcripts. The lexico-syntactic features were the most useful for this task. The automatic punctuation (comma, period, question mark) of French and English speech-to-text data was studied in [8]. Their boosting-based model uses linguistic (based on word n-grams) and prosodic information and was tested under real world conditions.

Based on the state of the art of question detection, we apply multiple feature combinations on our French data. Several approaches are analyzed: creating a classifier with only prosodic features or one with only linguistic features or one that combines both linguistic and prosodic features. Moreover classifier evaluations are carried out using linguistic features stemming out, on the one hand, from manual transcriptions, and on the other hand, from automatic speech-to-text transcriptions.

The work presented in this paper is part of the RAPSODIE project which aims at studying, deepening and enriching the extraction of relevant speech infor-

mation, in order to support communication with deaf or hard of hearing people. The detection of sentence modality (questions versus statements) is therefore a key problem here, the deaf or hard of hearing people must be informed when a question is directed to them, as they should respond or ask for further clarifications.

The paper is organized as follows: section 2 is devoted to the description of the data and tools used in our experiments, section 3 provides a description of the features used for question detection, and section 4 analyzes the results.

2 Experimental setup

2.1 Textual data for training language models

Textual punctuated data is used for modeling the lexical and syntactic characteristics of questions and statements. The available data corresponds to more than 800 million words from the French Gigaword corpus [11]. Based on a vocabulary of 97K words, 89K questions and 16M statements were extracted from this corpus by filtering the sentences ending with a question mark, respectively with a dot. The lexical data was also annotated with part-of-speech (POS) tags; this provided the syntactic data.

Based on the lexical (word-based) data we learned two language models, one for questions and one for statements, with a shared lexicon of 97K words. These language models have the purpose of representing the main word sequences that occur in a question rather than a statement (like for example in French: “est-ce que ...”, “qu’est-ce que ...”, etc).

Based on the syntactic (POS-based) data we learned two other language models, one for questions and one for statements, with a shared lexicon of 36 POS tags. These language models have the purpose of representing the main syntactic sequences that occur in a question rather than a statement (like for example in French the verb-pronoun inversions: “regardez vous ...”, “pourrait on ...”, “fallait il ...”, etc).

Table 1 describes the resulting 3-gram language models based on questions and statements, when using word-based sentences or POS-based sentences.

Table 1. Number of 3-grams in the language models computed over questions and statements

Language model	word-based	POS-based
questions	718K	9K
statements	68M	16K

2.2 Speech and textual data for modality detection

The speech corpora used to train and evaluate the modality detection classifiers (questions versus statements) come from the ESTER2 [12] and ETAPE [13] eval-

uation campaigns, and from the EPAC [14] project. The ESTER2 and EPAC data are French broadcast news collected from various radio channels (prepared speech and interviews). The ETAPE data correspond to debates collected from various radio and TV channels (spontaneous speech). These corpora were manually transcribed and punctuated (the segmentation of speech into sentences is therefore already given).

The set of questions and statements were extracted from these corpora by filtering the sentences ending with a question mark and respectively with a dot. The training sets of ESTER2, EPAC and ETAPE corpora are used to train the question detection classifiers; the development and test sets of the ESTER2 and ETAPE corpora are used to evaluate them.

The speech training data set contains 10K questions and 98K statements. However, binary classifiers do not work well when trained with imbalanced data sets: new instances are likely to be classified as the class that has more training samples. In order to avoid this overfitting problem, we chose to resample the data set by keeping all questions and randomly extracting subsets of statements of the same size (ten different training data sets are considered based on the different random lists of statements). In the 'Experiments and results' section we present only the average performance (with the associated standard deviation) over all ten training data sets.

Table 2 gives more details on the number of questions and statements used in our experiments.

Table 2. Description of the data used in our experiments

Data	**# questions**	**# statements**
Training data	10077	10077
Evaluation data	831	7005

2.3 Configuration

The SRILM tools [15] were used to train the statistical language models. The TreeTagger software [16] was used to annotate the transcriptions with POS tags.

The WEKA software [17] was used to train and evaluate 5 question detection classifiers:

- logistic regression (LR) [18],
- C4.5 decision tree (J48) [19],
- rule learner (JRip - Repeated Incremental Pruning to Produce Error Reduction) [20],
- sequential minimal optimization algorithm for training a support vector classifier (SMO) [21],
- neural network using backpropagation to classify instances (MP - Multilayer Perceptron) [22].

The values of F0 in semitones and of the energy are computed every 10 ms from the speech signal using the ETSI/AURORA acoustic analysis [23].

The forced speech-text alignment is carried out with the Sphinx3 tools [24]. This provides the speech segmentation into phones and words, which is then used to compute the sound durations, as well as to obtain the location and the duration of pauses. As the speech signal quality is rather good, it can be assumed that the segmentation is obtained without major problems.

The pronunciation variants were extracted from the BDLEX lexicon [25] and from in-house pronunciation lexicons, when available. For the missing words, the pronunciation variants were automatically obtained using JMM-based and CRFbased Grapheme-to-Phoneme converters [26].

The Sphinx3 tools were also used to train the phonetic acoustic models and to decode the audio signals. More information on the large-vocabulary decoding system used in our experiments and its associated lexicon can be found in [27, 28].

3 Features for question detection

3.1 Linguistic features

Three linguistic features were used to distinguish questions from statements:

– Two log-likelihood ratios (*lexLLR, synLLR*)

Two of our linguistic features are represented by the difference between the log-likelihood of the sentence with respect to the 'question' language model and the log-likelihood of the sentence with respect to the 'statement' language model (as done in [3] for Chinese). Computed as:

$$\mathrm{LLR}(\mathrm{sentence}) = \mathrm{Log}\left(\frac{\mathrm{P}(\mathrm{sentence}|\mathrm{questionLM})}{\mathrm{P}(\mathrm{sentence}|\mathrm{statementLM})}\right) \quad (1)$$

A sentence having a positive *LLR* value is likely to be a question. And vice-versa, a sentence having a negative *LLR* value is likely to be a statement.

To compute the lexical log-likelihood ratio (*lexLLR*) of a sentence we apply the lexical language models (of questions and statements) on its sequence of words.

To compute the syntactic log-likelihood ratio (*synLLR*) of a sentence we apply the syntactic language models (of questions and statements) on its sequence of POS tags.

– Presence of discriminative interrogative patterns (*iP*)

This feature indicates the presence (1) or absence (0) of some discriminative interrogative words or expressions. A sentence having an interrogative pattern is likely to be a question.

A list of sequential patterns was thus extracted from the Gigaword questions transcript with a modified version of the PrefixSpan software [29] that considers only consecutive patterns. Their frequencies were then compared between the Gigaword questions and statements transcripts: those with similar frequencies were removed. The patterns with no interrogative meaning were also removed.

The final list of discriminative interrogative patterns is: {quel, quelle, quels, quelles, comment, combien, pourquoi, est ce que, est ce qu', qu' est ce, qu' est ce que, qu' est ce qu'}, corresponding to {what, which, how, how much, why, ...}.

3.2 Prosodic features

The prosodic features include duration, energy and pitch belonging to the last prosodic group of the sentence. Prosodic groups are determined according to linguistic information (for grouping grammatical words with corresponding lexical words) and further processing that relies on prosodic information as described in [30]. Ten prosodic features were considered in order to distinguish questions from statements. Five are associated to the last syllable of the sentence, and five other are computed on the ending part of the sentence.

The duration of the last vowel is computed from the phonetic segmentation that results from the forced alignment. Its energy corresponds to the mean value calculated over all the frames of the vowel segment. The vowel energy and the vowel duration are then normalized with respect to local mean values computed on non-stressed vowels of the current breath group (speech segment delimited by pauses). In practice we used the vowels that are not in a word final position. The F0 slope is calculated by linear regression on the speech frames corresponding to the vowel. In addition to the slope, we calculate also, for the vowel, the delta of F0 movement with respect to the preceding vowel. The fifth parameter is the product of the F0 slope by the square of the vowel duration (this is inspired from the glissando threshold). Other, more global, prosodic parameters are computed on the longest F0 slope that ends in the last syllable of the sentence. Starting from the last syllable, we go back in time up to detecting an inversion of the F0 slope. We then compute parameters on this longest final F0 slope: the F0 slope itself (determined by linear regression), the length of this longest slope, the total F0 variation between the beginning and the end of the slope, and also the product of the slope by the square of the duration. One last prosodic parameter is used, which corresponds to the F0 level at the end of the sentence, expressed as the percentage of the speaker F0 range (0 corresponding to the lowest F0 value for the speaker, 100 corresponding to the highest F0 value for the speaker).

4 Experiments and results

The classifiers evaluated in our experiments (logistic regression, J48 decision tree, JRip rule learner, SMO sequential minimal optimization algorithm, neural network MP) gave similar results. Thus, only the results obtained with the classifier J48 are presented below.

The classifier evaluations are carried out using features stemming out from:

– automatic transcriptions (obtained with a large vocabulary speech recognizer) - to study the performance under real conditions
– manual transcriptions - to study the classifier's maximum performance, obtainable only in ideal conditions (i.e. with perfect transcripts).

The performance obtained on our imbalanced test data set (831 questions and 7005 statements) is evaluated by the harmonic mean between the ratio of correctly classified questions and the ratio of correctly classified statements, computed as:

$$\mathrm{H} = 2 * \left(\frac{\mathrm{ccQuestions} * \mathrm{ccStatements}}{\mathrm{ccQuestions} + \mathrm{ccStatements}} \right) \tag{2}$$

where "cc" is an acronym for "correctly classified". This value allows us to estimate the global performance of our classifier, given that the performances achieved on questions and on statements are equally important.

4.1 Prosodic features

The evaluated combinations of prosodic features are:

– the last F0 level ($lastF0level$)
– the 5 features computed over the last syllable ($lastSyl$),
– the 5 features computed over the last syllable plus the last F0 level ($lastSyl+lastF0level$),
– the 5 features computed over the ending part of the utterance ($lastPart$),
– the 6 features related to slope measurements ($slope$),
– all 10 features ($Prosodic$).

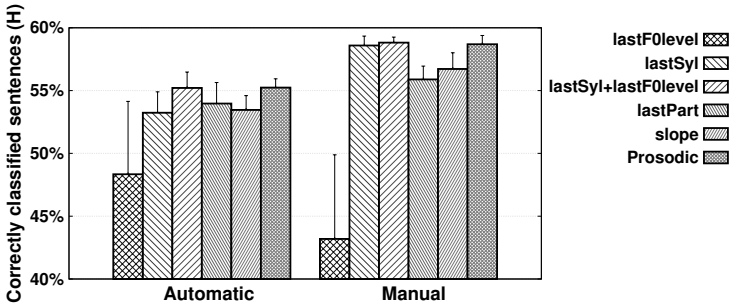

Fig. 1. Analysis of the average performance obtained when using different prosodic feature combinations on automatic and manual transcriptions

Figure 1 shows the average performance obtained with different prosodic feature combinations on automatic and manual transcriptions. The most important

prosodic features are those computed over the last syllable of the utterance in combination with the last F0 level ($lastSyl + lastF0level$). Combining all 10 prosodic features ($Prosodic$) does not deteriorate this performance: they are all considered to be useful and kept in the following experiments. The performance loss between manual and automatic transcriptions (of about 3%) is due to recognition errors and to the automatic word (phone) segmentation.

4.2 Linguistic features

The evaluated combinations of linguistic features are:

- the lexical log-likelihood ratio ($lexLLR$)
- the syntactic log-likelihood ratio ($synLLR$),
- the lexical log-likelihood ratio plus the presence of discriminative interrogative patterns ($lexLLR + iP$),
- the syntactic log-likelihood ratio plus the presence of discriminative interrogative patterns ($synLLR + iP$),
- both log-likelihood ratios ($lexLLR + synLLR$),
- all 3 features ($Linguistic$).

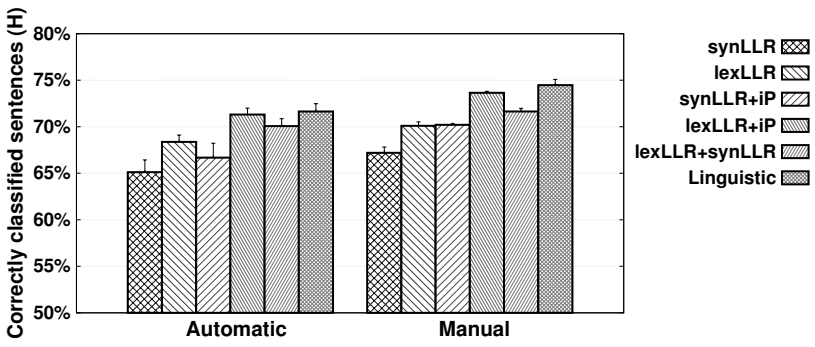

Fig. 2. Analysis of the average performance obtained when using different linguistic feature combinations on automatic and manual transcriptions

Figure 2 shows the average performance obtained with different linguistic feature combinations on automatic and manual transcriptions and it proves the importance of using and combining all of them. The most important linguistic feature is the lexical log-likelihood ratio ($lexLLR$). Combining it with the presence of discriminative interrogative patterns ($lexLLR + iP$) is more useful than combining it with the syntactic log-likelihood ratio ($lexLLR + synLLR$). The performance loss between manual and automatic transcriptions of the combined set of interrogative patterns and syntactic log-likelihood ratio ($synLLR + IP$)

is bigger than the one achieved by the *lexLLR* feature, which means that they are less tolerant to recognition errors. However, the combination of all three features (*Linguistic*) improves the classification performance, especially when dealing with correct transcriptions.

4.3 Combined prosodic-linguistic features

Table 3 shows the average performance (harmonic mean H, along with the ratios of correctly classified questions and correctly classified statements) obtained with the prosodic, linguistic and combined features, when applied on automatic speech-to-text transcriptions and on manual transcriptions. It can be easily observed that the linguistic classifiers outperform the prosodic classifiers. The performance obtained with the linguistic classifiers when applied on the automatic transcriptions and on the manual transcriptions differs by about 3% absolute, due to recognition errors (22% word error rate on Ester and 28% on Etape) and most likely to the misrecognition of the interrogative words. The combination of linguistic and prosodic features does not provide any improvement on manual transcripts and provides only a slight improvement on automatic transcription.

Table 3. Average performance (harmonic mean H, along with the ratios of correctly classified questions and correctly classified statements respectively) obtained on automatic and manual transcriptions, for prosodic features alone, linguistic features alone and with a combination of prosodic and linguistic features

Transcripts	**Prosodic**	**Linguistic**	**Combined**
automatic	55.24 (51.71; 60.23)	71.64 (66.62; 77.77)	72.21 (69.55; 75.25)
manual	58.69 (57.97; 59.55)	74.47 (71.57; 77.93)	74.26 (75.18; 73.46)

Table 4 gives more detailed results obtained with the combined prosodic-linguistic classifier on the manual transcriptions, when trained on a single random training set. 627 out of 831 questions were correctly classified as questions (ccQ=75.45%) and 5047 out of 7005 statements were correctly classified as statements (ccS=72.05%). The harmonic average performance is here H=73.71%.

Table 4. Confusion matrix between questions and statements obtained on manual transcriptions with the combined prosodic-linguistic classifier, when trained on a single random training set

	classified as **Question**	classified as **Statement**	
Question	627	204	**ccQuestions=75.45%**
Statement	1958	5047	**ccStatements=72.05%**

4.4 Combined outputs

A final experiment consisted in combining the outputs of all five classifiers (when using all 13 prosodic-linguistic features). Each classifier makes a class prediction (question or statement) on each utterance of the test data set. The final decision is made by a majority vote: if most of the classifiers (in this case at least 3) assign the utterance to class "question", than the utterance is assigned to class "question"; if not, than the utterance is assigned to class "statement".

Table 5 shows the average performance (H) obtained with all five classifiers separately, and with their combination (by majority vote). The majority vote and the 5 classifiers have similar performances, thus confirming that the 5 classifiers are likely to agree on the class predictions.

Table 5. Average performance (H) obtained with all 5 classifiers and with their combination (by majority vote) on manual and on automatic transcriptions

	LR	J48	JRip	SMO	MP	combination
Automatic	72.04	72.21	72.81	69.56	72.07	72.66
Manual	73.34	74.26	74.12	72.09	74.33	74.91

5 Conclusions

This paper analyzed the impact of linguistic features, prosodic features and combined linguistic-prosodic features when developing an automatic question detector. The context of this work is to support the communication with deaf or hard of hearing people, which requires an automatic detection of questions in order to inform them when a question is directed to them. The experiments were carried out using three French speech corpora: ETAPE, EPAC and ESTER2.

Different types of classifiers (logistic regression, decision tree, rule learner, sequential minimal optimization algorithm, neural network) were evaluated, but they all give similar results.

The prosodic classifier (based on 10 prosodic features) has a poor performance: it hardly exceeds 55% of correctly classified sentences. The most important prosodic features are those computed over the last syllable, in combination with the last F0 level.

The linguistic classifier (based on 3 linguistic features) provides by far better results: 72% when it is applied on ASR transcriptions (with perfect sentence boundaries) versus 74% when it is applied on reference manual transcripts. The most important linguistic feature is the lexical log-likelihood ratio (computed with respect to word-based language models).

The combination of prosodic and linguistic features does not provide any improvement on manual transcripts, but it provides a slight improvement on automatic transcription.

Future work will investigate further prosodic and linguistic features; confidence measures will also be considered in the computation of the linguistic features.

Acknowledgements The work presented in this article is part of the RAPSODIE project, and has received support from the "Conseil Régional de Lorraine" and from the "Région Lorraine" (FEDER) (http://erocca.com/rapsodie).

References

1. Jurafsky, D., Bates, R., Coccaro, N., Martin, R., Meteer, M., Ries, K., Shriberg, E., Stolcke, A., Taylor, P., Van Ess-Dykema, C.: Automatic detection of discourse structure for speech recognition and understanding. In: IEEE Workshop on Automatic Speech Recognition and Understanding. pp. 88–95 (1997)
2. Kral, P., Kleckova, J., Cerisara, C.: Sentence modality recognition in French based on prosody. In: International Conference on Enformatika, Systems Sciences and Engineering - ESSE 2005. vol. 8, pp. 185–188 (2005)
3. Yuan, J., Jurafsky, D.: Detection of questions in Chinese conversational speech. In: IEEE Workshop on Automatic Speech Recognition and Understanding. pp. 47–52 (2005)
4. Quang, V.M., Castelli, E., Yen, P.N.: A decision tree-based method for speech processing: question sentence detection. In: Proceedings of the Third international conference on Fuzzy Systems and Knowledge Discovery. pp. 1205–1212 (2006)
5. Quang, V.M., Besacier, L., Castelli, E.: Automatic question detection: prosodic-lexical features and crosslingual experiments. In: Proceedings of Interspeech. pp. 2257–2260 (2007)
6. Khan, O., Al-Khatib, W.G., Cheded, L.: A preliminary study of prosody-based detection of questions in Arabic speech monologues. Arabian Journal for Science and Engineering 35(2C), 167–181 (2010)
7. Margolis, A., Ostendorf, M.: Question detection in spoken conversations using textual conversations. In: Association for Computational Linguistics. pp. 118–124 (2011)
8. Kolar, J., Lamel, L.: Development and evaluation of automatic punctuation for French and English speech-to-text. In: Proceedings of Interspeech (2012)
9. Liscombe, J., Venditti, J.J., Hirschberg, J.: Detecting question-bearing turns in spoken tutorial dialogues. In: Proceedings of Interspeech (2006)
10. Boakye, K., Favre, B., Hakkani-Tur, D.: Any questions? automatic question detection in meetings. In: IEEE Workshop on Automatic Speech Recognition and Understanding. pp. 485–489 (2009)
11. Mendonça, A., Graff, D., DiPersio, D.: French Gigaword third edition. In: Proceedings of the Linguistic Data Consortium (2011)
12. Galliano, S., Gravier, G., Chaubard, L.: The ESTER 2 evaluation campaign for rich transcription of French broadcasts. In: Proceedings of Interspeech (2009)
13. Gravier, G., Adda, G., Paulson, N., Carré, M., Giraudel, A., Galibert, O.: The ETAPE corpus for the evaluation of speech-based TV content processing in the French language. In: Proceedings of the International Conference on Language Resources, Evaluation and Corpora (LREC) (2012)

14. Estève, Y., Bazillon, T., Antoine, J.Y., Béchet, F., Farinas, J.: The EPAC corpus: manual and automatic annotations of conversational speech in French broadcast news. In: Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC) (2010)
15. Stolcke, A.: SRILM an extensible language modeling toolkit. In: Conference on Spoken Language Processing (2002)
16. Schmid, H.: Probabilistic part-of-speech tagging using decision trees. In: Proceedings of the International Conference on New Methods in Language Processing. pp. 44–49 (1994)
17. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The WEKA data mining software: An update. SIGKDD Explorations 11(1), 10–18 (2009)
18. le Cessie, S., van Houwelingen, J.: Ridge estimators in logistic regression. Applied Statistics 41(1), 191–201 (1992)
19. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers Inc. (1993)
20. Cohen, W.: Fast effective rule induction. In: Proceedings of the Twelfth International Conference on Machine Learning. pp. 115–123 (1995)
21. Keerthi, S., Shevade, S., Bhattacharyya, C., Murthy, K.: Improvements to Platt's SMO Algorithm for SVM Classifier Design. Neural Computation 13(3), 637–649 (2001)
22. Ruck, D.W., Rogers, S.K., Kabrisky, M., Oxley, M.E., Suter, B.W.: The multilayer perceptron as an approximation to a Bayes optimal discriminant function. IEEE Transactions on Neural Networks 1(4), 296–298 (1990)
23. ETSI ES 202 212: Speech processing, transmission and quality aspects (STQ); distributed speech recognition; extended advanced front-end feature extraction algorithm; compression algorithms. ETSI ES (2005)
24. Placeway, P., Chen, S., Eskenazi, M., Jain, U., Parikh, V., Raj, B., Ravishankar, M., Rosenfeld, R., Seymore, K., Siegler, M., Stern, R., Thayer, E.: The 1996 Hub-4 Sphinx-3 System. In: DARPA Speech Recognition Workshop (1996)
25. de Calmès, M., Pérennou, G.: BDLEX : a lexicon for spoken and written French. In: Proceedings of the International Conference on Language Resources and Evaluation (LREC). pp. 1129–1136 (1998)
26. Jouvet, D., Fohr, D., Illina, I.: Evaluating grapheme-to-phoneme converters in automatic speech recognition context. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4821–4824 (2012)
27. Jouvet, D., Fohr, D.: Combining forward-based and backward-based decoders for improved speech recognition performance. In: Proceedings of Interspeech (2013)
28. Jouvet, D., Langlois, D.: A Machine Learning Based Approach for Vocabulary Selection for Speech Transcription. In: TSD - 16th International Conference on Text, Speech and Dialogue - 2013. vol. 8082, pp. 60–67 (2013)
29. Pei, J., Han, J., Mortazavi-asl, B., Pinto, H., Chen, Q., Dayal, U., Hsu, M.C.: PrefixSpan: Mining sequential patterns efficiently by prefix-projected pattern growth. In: International Conference on Data Engineering. pp. 215–224 (2001)
30. Bartkova, K., Jouvet, D.: Automatic detection of the prosodic structures of speech utterances. In: Speech and Computer, vol. 8113, pp. 1–8. Springer (2013)