Interspeech 2018
2-6 September 2018, Hyderabad

Classification of Correction Turns in Multilingual Dialogue Corpus

Ivan Kraljevski, Diane Hirschfeld

voice INTER connect GmbH, Dresden, Germany

{ivan.kraljevski,diane.hirschfeld}@voiceinterconnect.de

Abstract

This paper presents a multiclass classification of correction dialog turns using machine learning. The classes are determined by the type of the introduced recognition errors while performing WOz trials and creating the multilingual corpus. Three datasets were obtained using different sets of acoustic-prosodic features on the multilingual dialogue corpus.

The classification experiments were done using different machine learning paradigms: Decision Trees, Support Vector Machines and Deep Learning. After careful experiments setup and optimization on the hyper-parameter space, the obtained classification results were analyzed and compared in the terms of accuracy, precision, recall and F1 score. The achieved results are comparable with those obtained in similar experiments on different tasks and speech databases.

Index Terms: multiclass classification, machine learning, multilingual dialogue corpus

1. Introduction

In Spoken Dialog Systems (SDS), Automatic Speech Recognition (ASR) and Natural Language Understanding (NLU) are challenging tasks and errors are still ultimately unavoidable. In reality, there is no ideal speech interface and problems in human-computer conversation mostly arise in cases of miscommunication between the interacting sides, regardless the cause: miss- or non-recognition (ASR), misunderstanding (NLU), inappropriate prompting or wrong dialog context (Dialog Manager-DM). Therefore, it is of great importance to implement an appropriate recovery and error handling strategy, as close as possible to the way humans would react in such situations. This is only possible if the system is capable of being aware of problematic communication.

Many research groups are dealing with the topic of prediction, detection and reduction of miscommunication in Spoken Dialog Systems. In [1], the data-driven approach for detecting instances of miscommunication is described. Handcrafted rule-based methods are presented in [2-3], Bayesian networks were used in [4-5], discriminative models in [6], and Long Short-Term Memory Neural Networks in [7].

The authors in [8] proposed a system which integrates an error correction detection module with a modified dialogue strategy. In the study [9], a machine-learning approach employed automatically derived prosodic features, the speech recognition process, experimental conditions and the dialogue history to identify user corrections of speech recognition errors. An error handling strategy based on dynamically created correction grammars for recognizing correction sentences is described in [10]. Other research studies used different sources of information to detect problematic turns, in [11] the authors used information from the language model to train an ANN that detected mis-recognized words and out-of-scope phrases, while in [12], the authors combined information from the speech recognizer, parser, and the dialogue manager.

The speaking style changes associated with correction dialogue acts are characterized by distinctive prosodic features mostly correlated with hyperarticulated speech. Thus, hyperarticulation can be used as a clue in order to identify problematic turns. Using prosodic features for recognizing and classifying dialogue acts was investigated in [13]. In [14] the duration, pause, and pitch features were employed to train a decision tree classifier, which was extended and integrated with recognizer confidence scores for further improvements in the detection of corrections [15].

The authors in [16] observed that human speech during error resolutions shifts to become lengthier and more clearly articulated. A similar study presented in [17] shows that English speaker's utterances of correction and non-correction dialogue acts differ prosodically in ways consistent with hyperarticulated speech. They defined it as: "slower and louder speech with wider pitch excursion and more internal silence", similar findings were reported for German speech data in [18]. Hyperarticulation detection is a challenging task for the humans and for the computers. The users have different speaking styles which make it challenging to actually see that they are hyperarticulating. Classification of a single utterance regardless the previous one could lead to poor classification performance. The studies [19] and [20] avoid the problem by considering a pair of the user utterances spoken in sequence.

In our previous work [21], cross-linguistic differences related to hyperarticulated speech in correction dialogue acts were investigated. It was confirmed that there are distinctive prosodic features across 9 different languages associated with hyperarticulated speech. In general, the speakers raised their voice (pitch and intensity) in the case of reacting on the request to repeat the last utterance (deletions) but they did the opposite in the case of insertions, mostly confused by the sudden and unexpected system confirmation. The speech rate (including the pauses and hesitations) was slower in misrecognition clarifications (substitutions).

While there are many studies successfully dealing with automatic detection of correction dialog acts, very few like [9], attempt to classifying them in more elementary categories according the cause (non-recognition, non-understanding, misunderstanding, etc). The aim of this paper is to investigate the possibility of analysis and classification of correction dialogue acts in multilingual corpus, further than detecting presence-absence of distinctive prosody features indicating miscommunications.

10.21437/Interspeech.2018-1348

2. Material and methodology

2.1. Speech database

For the experiments, we used the speech database, collected in WOz tests involving participants of 13 different languages interacting with a smart-home system [22]. Such parallel multilingual corpus is a solid basis to perform investigations on the behavioral patterns of native speakers, from the linguistic as well as para-linguistic aspects.

The speech database consist of approximately 4500 orthographically transcribed scenarios with a total duration of 125 hours. The following languages were covered (abbreviation and number of participants in brackets): English (EN:40), German (DE:40), French (FR:23), Spanish (ES:27), Italian (IT:19), Dutch (NL:15), Finnish (FI:7), Norwegian (NO:7), Swedish (SE:6), Danish (DA:8), Russian (RU:20), Turkish (TR:20) and Mandarin Chinese (CN:19). During the sessions, the wizard triggered spoken dialogue acts and device functions to simulate a perfect dialogue system. Miscommunication was simulated by introducing embedded error speech prompts, categorized as:

- Substitutions: wrongly recognized parameters;
- Insertions: confirmation of non-uttered sentence;
- Deletions: request to repeat the last sentence.

The maximum number of introduced errors (around 20%) in a session was estimated over the number of the required parameters (options, entries) per scenario, including occasional system rejections and repetitions. They were not triggered automatically and not all the planned error prompts are played since the actual dialogue flow never reached intended states.

The total distribution of the paired dialogue turns on all languages is: deletions 35.20% (1182), insertions 8.04% (270) and substitutions 56.76% (1906). Figure 1 presents the introduced error distribution across languages. For some languages, there are differences in the count of insertions errors, because often the speakers were quite confused providing no answer that could be paired with the statement.

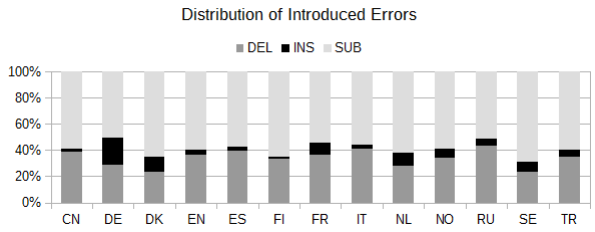

Figure 1: *Distribution of the Introduced Errors*

2.2. Data organization

Common datasets were compiled for all languages, based on the collected corpus and the time-stamped logs of the dialogue acts. We selected pair of utterances of "statement" and "correction" dialogue turns (in total 3026). The dataset was divided by randomly sampling into a training (80% or 2686 observations) and a test set (20% or 672 observations).

It has to be emphasized that the collected speech for the correction turns were only transcribed by orthography and not evaluated by any other speech characteristic. That means there are no annotations describing presence of hyperarticulated speech. This makes the problem of classifying the dialog correction acts into subcategories even more challenging and the performance probably will not reach those achieved on different databases.

2.3. Acoustic-prosodic features

We employed 3 different acoustic-prosodic feature extraction procedures for the paired turns. The methods produced features values for the "statement" turns, which were subtracted from those of the corresponding "correction" turns, producing datasets representing quantitative changes in acoustic-prosodic features over the complete sentence.

Such delta values are considered better suited for analysis, compensating speaker and environment specific influences [20]. Statistical analysis presented in [21] showed that, after Shapiro Wilk normality test on the delta features of the VIC dataset, regardless of the language, they are not represented with a normal distribution, as is usually true for a large sample count real data.

Non-parametric Wilcoxon test applied to all the delta values confirmed the presence of distinctive prosodic features, particularly related to slower speech, indicating hyperarticulation.

2.3.1. VIC features

The first, noted as VIC in the following text was built using the following Praat [23] scripts.

"Praat Script Syllable Nuclei v2" [24] was used for automatic detection of syllable nuclei in order to estimate the speech rate without the need of manual transcription. Peaks in intensity (dB) that are preceded and followed by dips in intensity are considered as potential syllable nuclei, while the peaks that are not voiced were discarded. The following measures were considered: speech rate (nsyll/speech-duration), articulation rate (nsyll/phonation-time) and average syllable duration (phonation-time/nsyll). Where *nsyll* is the number of syllables detected in either speech duration or phonation time.

"ProsodyPro 6beta" [25] was used for systematic analysis of the datasets to generate detailed discrete prosodic measurements suitable for statistical analysis: maximal *f0* (Hz), minimal *f0* (Hz), pitch excursion (semitones), averaged *f0* (Hz), averaged intensity (dB) and maximum *f0* velocity (semitone/s).

2.3.2. IS09 emotion features

In addition to VIC features, we used the standard feature set designed for emotion recognition: the Interspeech 2009 (IS09) emotion challenge feature set. It contains 384 features extracted from open source feature extraction toolkit openSMILE [26]. The influence of emotion on the articulation degree has been studied in [27], which give us the idea that the IS09 emotion feature set could be useful for analysis of hyperarticulated speech.

2.3.3. IS13 ComParE features

Apart from the smaller set, we employed also features which were used as a baseline in the Interspeech 2013 (IS13) ComParE Challenge [28]. The set contains 6373 features derived by processing of low-level descriptor (LLD) contours extracted by openSMILE. The LLD features include pitch

(fundamental frequency), intensity (energy), spectral, cepstral (MFCC), duration, voice quality (jitter, shimmer, and harmonics-to-noise ratio), spectral harmonicity, and psychoacoustic spectral sharpness.

This standard feature set has been successfully used for many computational paralinguistic tasks, including emotion recognition, native language detection, sincerity, etc.

3. Experiments and results

3.1. Experiments setup

In the experiments, we considered the detection of different types of correction turns with present hyperarticulated speech as a multiclass classification problem. The objective was to find out which approach for non-linear classification is best suited for the datasets. The datasets are characterized by small number of unbalanced classes and small amount of training data per class. In all of the experiments, we are using the R package for statistical computing [29].

The choice of an appropriate classification approach depends by a number of factors. In this case, particularly important are the: 1) tolerance of high dimensionality, 2) capability of exploiting a small dataset, and 3) handling of unbalanced classes.

At first, we performed non-linear classification with Decision Trees (DT) which were successfully applied in similar investigations [30]. Then, Random Forests (RF) [31], an approach based on decision trees, which has been proved successful in experiments using similar data [32].

Followed by commonly used Support Vector Machine (SVM) classification which seems well suited when applied on the OpenSMILE derived acoustic-prosodic features.

At the end, we assessed the usability of Deep Neural Networks (DNN) in comparison with the other methods, considering the limitations of a rather small number of observations, inconsistent data set and a large feature dimensions.

3.2. Classification methodology

In the classification tests with Decision Trees, the class weights or prior probability were applied correspondingly to overcome the problem of unbalanced datasets. For all experiments, we used 5-fold Cross Validation (CV) on the train set and measured the mean and the standard deviation on the original test set across the folds for unweighted: accuracy, precision, recall and F1 score. To ensure the repeatability of the experiments, we kept the same division for the training and test set, as well as the validation folds.

The variable importance, ranked by the mean decrease in Gini coefficients was also investigated. For the VIC dataset, the most important factors correspond with the significant ones found by Linear Mixed Model analysis in [20]. However, there was no large difference in the mean decrease in Gini coefficients in order to achieve any improvements by omitting some of the factors. We assumed that larger feature sets could benefit from excluding most of the factors which are not significant predictors. By investigating the slope of the mean decrease in Gini coefficient function we choose the approximate cut-off points by looking for a larger difference between the factors.

There is the risk of having either too few variables (which could not provide proper separation) or too many (which will over-explain the differences). For the IS09 dataset, the feature dimension was reduced from 384 to 46 and for the IS13 set from 6373 to 383 variables.

3.2.1. Decision Trees

We assumed linear dependencies in the data set, consecutively we employed Recursive Partitioning (RP) decision trees to build the first classification model. The trees were pruned to the optimal value of the complexity parameter.

For the second, Random Forest (RF) model, we explored different values for the number of trees (ranged from 2 to 512), as well as the maximum number for nodes (100 to 1400). It was observed, that for all datasets after increasing the number of trees as well the number of nodes, the classification performance on the test set did not increased further. That indicates over-fitting of the model, as pointed out also in [33].

3.2.2. Support Vector Machines

Although originally developed for binary classification, SVMs [34] are widely used also in multiclass recognition tasks. In order to achieve acceptable results, correct choice of kernel parameters is very important. Before the results can be trusted, an extensive search has be conducted on the hyper-parameter ranges to find the most optimal values.

To train our SVMs, we took advantage of the R interface to the well known LIBSVM library [35]. The Radial Kernel Function (RBF) was chosen because of its good general performance and the SVM was tuned over a range of the cost (10^{-4} to 10^{1}) and the gamma (10^{-9} to 10^{1}) parameters.

3.2.3. Deep learning models

In this section, the usability of Deep Learning paradigm for classification of correction turns was explored. DNN models (Multi-Layer Perceptron) were trained over the feature sets used in the previous experiments. The same sequence of experiments is performed on the "Full" and the "Selected" features. The R interface to Keras [36], the neural network API was employed, with the Tensorflow [37], as the back-end.

A grid search was performed over the hyper-parameter space, to get the most optimal values for the number of layers and nodes per layer. The dropout ratio was chosen to 0.495 in order to reduce the risk of over-fitting [30], which is emphasized in the case of a small amount of data and a larger number of features. The topology consisted of fully connected layers with equal number of hidden units and the ReLU activation function. The output layer has softmax activation and three output nodes corresponding to the target classes. During training, the categorical cross-entropy was used as a loss function, the output of each layer was normalized using batch normalization and passed through a dropout layer.

The models were trained using Adam [38] stochastic optimization which is well suited for tasks that are large in terms of data and/or parameters. The learning rate was set to 10^{-4} and the decay rate to 10^{-6} with batch size of 128.

The maximum number of epochs was set to 50 with the condition of 10 epochs with no improvement after which the training was stopped.

4. Discussion

Figure 2 presents the achieved results in terms of unweighted accuracy (UAR) and F1 score, along with the standard deviation, for the all three datasets with the original number of features (the "Full" datasets).

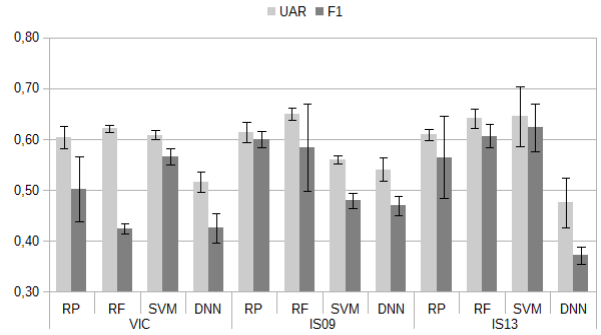

Figure 2. *UAR and F1 for the "Full" datasets*

Figure 3 shows the comparison of the F1 score between the datasets with "Full" and the "Selected" features.

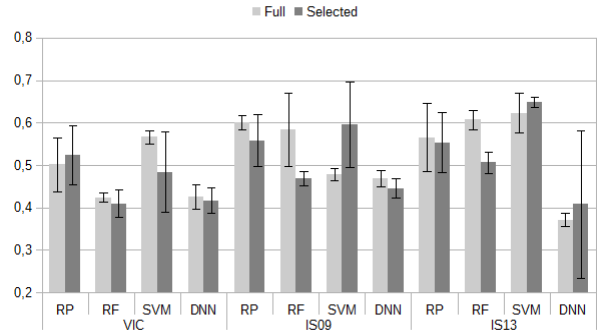

Figure 3. *F1 score of "Full" and "Selected" datasets*

Since the classes in the datasets are unbalanced, Weighted Guess Classifier was used as a baseline, with calculated accuracy of 0.452.

The OpenSMILE feature sets IS09 and IS13 provided in most of the cases better classification. The best performing model achieved an accuracy of 0.689 ± 0.015, precision 0.643 ± 0.018, recall 0.661 ± 0.024 and the F1 score of 0.649 ± 0.013 and it used SVM training with "Selected" IS13 dataset. The VIC dataset has the smallest number of features and that was the reason of worse performance in comparison with the other two. The best DNN model yielded an F1 score of 0.469 on the IS09 openSMILE "Full" dataset. All models provided results well above those of the baseline classifier, with the maximum improvement of 31.6%.

It is obvious from the results that the Deep Learning approach did not fulfill the expectations, the models could not reach the classification performance of the decision trees and SVM classifiers. The reasons are the relatively small amount of available data, in contrast with large number of features, the unbalanced class distribution and the non-consistent content of the classes. Namely, the classes were determined during the WOz experiments by introducing different types of errors (deletions, insertions and substitutions) and not by human annotations of the dialogue turns.

In general, the machine learning algorithms benefit from more features and selection of better predictors. This was confirmed after reducing variables to the most important ones, according the Gini index, where on the IS09 and IS13 datasets this provided improvements in classification results with SVM models. The multilingual nature of the data and the different speakers were not an influencing factor because the analysis was done on a pairwise dialog turns.

If we omit the insertion error class as the most inconsistent, the task is transformed to a binary classification problem. Then the best performing DNN topology trained on the IS09 "Full" dataset achieved: accuracy 0.630 ± 0.034, precision 0.620 ± 0.040, recall 0.617 ± 0.039, and F1 0.617 ± 0.038. This is improvement in accuracy of 10.9% compared for to the corresponding baseline classifier. The results are also comparable to those obtained in similar studies in case of binary classification [30].

The classification accuracy could be improved by re-annotating the correction turns of the multilingual corpora by human experts. That is, in general, challenging and difficult task since subtle differences in acoustic-prosodic characteristics should be observed and there would be substantial disagreement between the human annotators.

5. Conclusions

In this paper, we tackled the problem of multiclass classification of correction dialog turns exploiting acoustic-prosody features and machine learning. Many similar studies are dealing with classification of para-linguistic aspects in dialog turns, most of them as binary classification tasks, except in the cases, where adequate amount of data is available. The target classes in our experiments were determined during acquisition of the multilingual dialog corpus in WOz sessions, by the type of the introduced errors (deletions, substitutions and insertions). Three datasets were obtained from the corpus using different sets of acoustic-prosodic features, named as VIC, and the OpenSMILE IS09 and IS13.

For the classification experiments, Decision Trees, Support Vector Machines and Deep Learning machine learning paradigms were used. The achieved results were analyzed and compared in terms of unweighted accuracy, precision, recall and F1 score. The best performing model is based on the SVM approach and used the "Selected" IS13 OpenSMILE dataset. DNN models did not perform well enough on any of the datasets due to the relatively small amount of observations and larger number of features. When the task was reformulated as binary classification (deletions and substitutions errors) the DNN model provided results comparable with those obtained in similar tasks and on different speech databases.

In the future, we plan to re-evaluate the presented approaches on the multilingual corpus after improving the annotations regarding the target classes in the same time expecting improvements in classification performance.

6. Acknowledgments

The authors are thankful to Ingo Siegert for his comments on an earlier version of this paper.

7. References

[1] R. Meena, G. Skantze, and J. Gustafson, "Automatic detection of miscommunication in spoken dialogue systems", *in Proc. of SIGdial*, 2015.

[2] S. Larsson and D.R. Traum, "Information state and dialogue management in the TRINDI dialogue move engine toolkit". *Natural language engineering.* 2000 Sep; 6(3-4):323-40.

[3] D. Bohus and A. Rudnicky, "The RavenClaw dialog management framework: architecture and systems". *Computer Speech & Language*, *23*(3), pp.332-361, 2009.

[4] T. Paek and E. Horvitz, "Conversation as action under uncertainty", *in Proceedings of the Sixteenth conference on Uncertainty in artificial intelligence* (pp. 455-464), 2000, Morgan Kaufmann Publishers Inc.

[5] J.D. Williams and S. Young, "Partially observable Markov decision processes for spoken dialog systems", *in Computer Speech & Language*, 2007, 21(2), pp. 393-422.

[6] D. Bohus and A. Rudnicky, "A k-hypotheses+ other belief updating model", *in Proc. of the AAAI Workshop on Statistical and Empirical Methods in Spoken Dialogue Systems,* 2006 (Vol. 62).

[7] K. Yoshino, T. Hiraoka, G. Neubig and S. Nakamura, "Dialogue State Tracking using Long Short Term Memory Neural Networks", *in Proceedings of the Seventh International Workshop on Spoken Dialog Systems (IWSDS)*, 2016, pp. 1-8

[8] I. Bulyko, K. Kirchhoff, M. Ostendorf, and J. Goldberg, "Error-correction detection and response generation in a spoken dialogue system", *in Speech Communication,* vol. 45, no. 3, pp. 271-288, 2005.

[9] J. Hirschberg, D. Litman, and M. Swerts, "Characterizing and predicting corrections in spoken dialogue systems", *in Comput. Linguist*, vol. 32, pp. 417-438, 2006.

[10] H. Sagawa, T. Mitamura, and E. Nyberg, "Correction grammars for error handling in a speech dialog system", *in HLT/NAACL*, *Boston,* 2004.

[11] R. San-Segundo, B. Pellom, W. Ward, and J. M. Pardo, "Confidence measures for dialogue management in the CU Communicator system", *in Proc. of ICASSP*, 2000.

[12] D. Bohus and A. I. Rudnicky, "Integrating Multiple Knowledge Sources for Utterance-Level Confidence Annotation in the CMU Communicator Spoken Dialog System", *in Technical Report CS-190*, 2002.

[13] E. Shriberg, A. Stolcke, D. Jurafsky, N. Coccaro, M. Meteer, R. Bates, P. Taylor, K. Ries, R. Martin and C. Van Ess-Dykema, "Can Prosody Aid the Automatic Classification of Dialog Acts in Conversational Speech?", *in Language and Speech*, 41:439-487, 1998

[14] G.-A. Levow, "Characterizing and recognizing spoken corrections in human-computer dialogue", *in Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*. Association for Computational Linguistics, 1998.

[15] J. Hirschberg, D. Litman, and M. Swerts, "Prosodic and other cues to speech recognition failures", *in Speech Communication, 43(1-2):155–175,* 2004.

[16] S. Oviatt, "Modeling hyperarticulate speech during human-computer error resolution", *in Proc. of the Int. Conference on Spoken Language Processing*, pp. 797-800, 1996

[17] M. Swerts, D.J. Litman, J. Hirschberg, "Corrections in spoken dialogue systems", *in INTERSPEECH,* pp. 615-618, 2000.

[18] H. Soltau and A. Waibel, "Specialized acoustic models for hyperarticulated speech", *in Acoustics, Speech, and Signal Processing,* Vol. 3, pp. 1779-1782, 2000.

[19] A. Fandrianto, and M. Eskenazi, "Prosodic entrainment in an information-driven dialog system", *in Thirteenth Annual Conference of the International Speech Communication Association,* 2012.

[20] R.G. Kulkarni, A. El Kholy, Z. Al Bawab, N. Alon, I. Zitouni, U. Ozertem, S. Chang, "Hyperarticulation detection in repetitive voice queries using pairwise comparison for improved speech recognition", *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, pp. 4985-4989, 2017.

[21] I. Kraljevski, D. Hirschfeld, "Hyperarticulation of Corrections in Multilingual Dialogue Systems", *in Proc. Interspeech 2017*, (2017): 2531-2535

[22] I. Wendler, A. Jatho, I. Kraljevski, M. Wenzel, "Nutzerzentrierter Entwurf von Multimodalen Bedien-konzepten", *28. Konferenz Elektronische Sprach-signalverarbeitung 2017,* Universität des Saarlandes, Saarbrücken, 15.–17. März 2017.

[23] P. Boersma. "Praat, a system for doing phonetics by computer", *in Glot International* 5:9/10, 341-345, 2001.

[24] N.H. De Jong and T. Wempe, "Praat script to detect syllable nuclei and measure speech rate automatically", *in Behavior research methods*, 41 (2), 385 – 390, 2009.

[25] Y. Xu, "ProsodyPro - A Tool for Large-scale Systematic Prosody Analysis", *in Proceedings of Tools and Resources for the Analysis of Speech Prosody* (TRASP 2013), Aix-en-Provence, France. 7-10, 2013.

[26] F. Eyben, M. Wöllmer and B. Schuller, "Opensmile: the Munich versatile and fast open-source audio feature extractor", *in Proceedings of the 18th ACM international conference on Multimedia*. *ACM,* pp. 1459–1462, 2010.

[27] G. Beller, N. Obin and X. Rodet, "Articulation degree as a prosodic dimension of expressive speech", *In Fourth International Conference on Speech Prosody*, 681-684, 2008.

[28] B. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, M. Mortillaro, "The INTERSPEECH 2013 computational paralinguistics challenge: social signals, conflict, emotion, autism", *in Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association*, Lyon, France, 2013.

[29] R Core Team (2016), "R: A language and environment for statistical computing", *R Foundation for Statistical Computing, Vienna, Austria*. URL https://www.R-project.org.

[30] M. Gideon, S. I. Levitan, K.Z. Lee and J. Hirschberg. "Hybrid Acoustic-Lexical Deep Learning Approach for Deception Detection." *in Proc. Interspeech 2017* : 1472-1476, 2017.

[31] L. Breiman: "Random Forests". *Machine learning 45*, no. 1 (2001): 5-32.

[32] E. Ribeiro, F. Batista, I. Trancoso, R. Ribeiro, D.M. de Matos, "Automatic Detection of Hyperarticulated Speech", *in International Conference on Advances in Speech and Language Technologies for Iberian Languages* 2016 Nov 23 (pp. 182-191). Springer, Cham.

[33] T.M. Oshiro, P.S. Perez and J.A. Baranauskas, "How many trees in a random forest?", *in International Workshop on Machine Learning and Data Mining in Pattern Recognition*, 2012, July, (pp. 154-168). Springer, Berlin, Heidelberg.

[34] V. Vapnik, "Statistical learning theory", New York: Wiley, 1998

[35] C.C. Chang and C.J. Lin, "LIBSVM: a library for support vector machines", *in ACM transactions on intelligent systems and technology (TIST)*. 2011 Apr 1;2(3):27.

[36] F. Chollet, "Keras: Deep learning library for Theano and Tensorflow". URL: https://keras. io/k. 2015;7:8.

[37] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard and M. Kudlur, "TensorFlow: A System for Large-Scale Machine Learning", *in OSDI*, Vol. 16, pp. 265-283 (2016).

[38] D.P Kingma and J. Ba, "Adam: A method for stochastic optimization", *in arXiv preprint arXiv:1412.6980* (2014).