ANALYSIS AND CATEGORIZATION OF CORRECTIONS IN MULTILINGUAL SPOKEN DIALOGUE SYSTEM

I. Kraljevski, D. Hirschfeld

voice INTER connect GmbH, Ammonstraße 35, D-01067 Dresden, Germany
{ivan.kraljevski, diane.hirschfeld}@voiceinterconnect.de

Abstract: Human-machine conversation imposes many challenges where communications errors are still ultimately unavoidable. It is of great importance to facilitate the detection and correction of miscommunication. A robust dialogue system has to be able to detect miscommunication and to apply appropriate recovery and error strategies. This is only possible if the system is capable of being aware of any problematic communication by analyzing and classifying correction dialogue acts.

The speaking style changes, associated with corrections, are characterized by distinctive prosodic features. They are mostly correlated with hyperarticulated speech, which can be used as a clue to identify problematic situations. In this paper we analyzed, categorized and detected distinctive acoustic-prosodic features of corrections on 13 different languages. The statistical analysis showed that there is a significant relationship to the language and the type of correction with the features related to hyperarticulated speech. In general, speakers raised their voice in the case of a request to repeat the last utterance, but they did the opposite in the case of insertions, also the speech rate was slower in misrecognition clarifications.

Additionally, we presented the results of classification experiments of corrections exploiting acoustic-prosody feature analysis in combination with machine learning. The datasets are characterized by a small number of unbalanced classes and a small amount of training data per class. Support Vector Machines and Artificial Neural Networks were employed for the multi-class and binary classification. The results were analyzed and compared in terms of unweighted accuracy, precision, recall, and F1 score.

1 Introduction

In Spoken Dialogue 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. 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.

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

The authors in [6] proposed a system which integrates an error correction detection module with a modified dialogue strategy. In the study [7], 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 [8]. Other research studies used different sources of information to detect problematic turns, in [9] the authors used information from the language model to train an ANN that detected misrecognized words and out-of-scope phrases.

The speaking style changes associated with correction dialogue acts are characterized by distinctive prosodic features mostly correlated with hyperarticulated speech, which can be used as a clue to identify problematic turns. Using prosodic features for recognizing and classifying dialogue acts was investigated in [10]. In [11] 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 [12].

The authors in [13] observed that human speech during error resolutions shifts to become lengthier and more clearly articulated. A similar study presented in [14] 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". Hyperarticulation detection is a challenging task for humans and for computers. The speakers have different speaking styles which make it challenging to actually see that they are hyperarticulating, therefore classification of a single utterance could lead to poor classification performance. The studies [15] and [26] avoid the problem by considering an adjacent pair of utterances.

While there are many research studies dealing with cross-linguistic prosodic differences [13-14], they are mostly done using a pair of languages and on a limited number of participants. Moreover, very few like [7], attempt to classify correction acts in more elementary categories according to the cause (non-recognition, non-understanding, misunderstanding, etc.).

The aim of this paper is to analyze, categorize and detect corrections by their distinctive acoustic-prosodic features. Additionally, employing machine learning techniques, like Support Vector Machines (SVMs) and Artificial Neural Networks (ANNs), in combination with acoustic-prosody feature analysis, should give the answer to whether a reliable classification is possible at all. By our best knowledge, there are no research studies dealing with simultaneous analysis and categorization of corrections by their characteristic features of hyperarticulated speech in several languages in parallel.

2 Speech Database

The multilingual speech database was collected by staging a series of Wizard-of-Oz (WOz) experiments. In a preparatory phase, an online questionnaire with a total of 870 participants was carried out in 13 languages [17]. In the implementation phase, 19 different user scenarios and their corresponding dialogues for control of smart home devices were designed. The scenarios were carefully designed to elicit spontaneous reactions and to trigger recovery behavior from the participants in case of miscommunication.

The WOz experiments were carried out for the following languages (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). The dialogues were translated and adapted for all languages while keeping the same meaning and the semantic structure whenever possible.

During the session, the wizard manually triggered speech dialogue acts and devices function to simulate a perfect dialogue system. Miscommunication was simulated by presenting speech prompts, which elicited different types of correction responses: *Substitutions* - reaction on wrongly recognized parameters, *Insertions* - reaction to confirmation of a non-uttered sentence; *Deletions* - reaction on request to repeat the last utterance.

The maximum number of eliciting prompts, (around 20% for a session), was estimated over the number of required parameters (options, entries) per scenario, including occasional system rejections and repetitions. However, not all of the planned correction eliciting prompts were played since the actual dialogue flow never reached some states they should be presented at. Approximately 4500 scenarios for all languages were fulfilled yielding audio recordings of 125 hours.

3 Data organization

3.1 Acoustic-prosodic features

We employed 2 different acoustic-prosodic feature extraction procedures for the selected adjacent "statement - correction" dialogue acts pairs.

3.1.1 "VIC" features

Two Praat [18] scripts were used to calculate 16 acoustic-prosodic features:

- "*Praat Script Syllable Nuclei v2"* [19], 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 to be potential syllable nuclei, while the peaks that are not voiced were discarded. The following measures were used: speech rate (nsyll/speech-duration), articulation rate (nsyll/phonation-time) and average syllable duration (phonation-time/nsyll). *nsyll* is the number of syllables detected in either speech duration or phonation time.
- "*ProsodyPro 6beta"* [20], 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).

3.1.2 "IS09 emotion" features

In addition to the VIC features, we used the feature set designed for emotion recognition: the Interspeech 2009 (note as IS09) emotion challenge feature set. It contains 384 features extracted by the open source toolkit openSMILE [21]. The features are obtained by applying 12 functionals to the low-level descriptors: zero-crossing rate, root mean square energy, cepstrum computed *f0*, voicing probability computed by autocorrelation function and Mel-Frequency cepstral coefficients 1-12, together with their first order delta regression coefficients. The influence of emotion on the articulation degree has been studied in [22], which makes the IS09 emotion feature suitable to be used for analysis and classification of hyper-articulated speech in corrections.

3.2 Delta values

We selected adjacent pairs of utterances, "statement" and "correction" dialogue acts, for both feature dataset, providing in total 3303 observations. The datasets were transformed by subtracting the features of the "statement" acts from the adjacent "corrections" acts, providing quantitative measure about how acoustic-prosody features are changed through both acts. Such delta features are considered better suited for analysis and classification, compensating different speakers and environmental conditions.

4 Exploratory statistics

4.1 Correction responses distributions

The total distribution of the elicited corrective responses in all languages is deletions 35.24% (1164), insertions 7.78% (257) and substitutions 56.98% (1882).

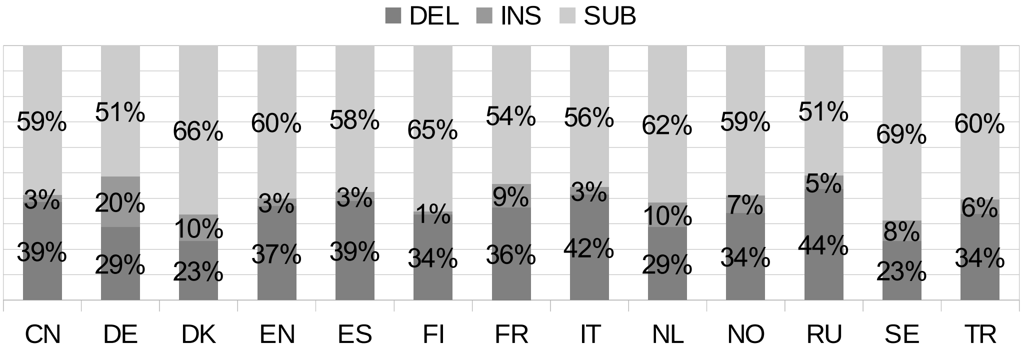

Figure 1 – Distribution of types of elicited corrections

In Figure 1 it can be seen that for some languages, there are differences in the count of insertion errors because often the speakers were quite confused providing no answer that could be paired with the “statement” act.

Table 1 – Correction response type examples and categorization

Type of Correction Response	Statement	System	Correction
Different, non-matching content (DIFF)	35 minutes	10 minutes	correction 35
Full, identical content (FULL)	35 minutes	10 minutes	35 minutes
Mixed, statement contained in correction (MIX)	35 minutes	10 minutes	no Vicky, 35 minutes
Partial, correction contained in statement (PART)	35 minutes	10 minutes	35

Table 1 shows an example of categorization of the transcribed corrections into four types of responses. The distribution of the response types is shown in Figure 2.

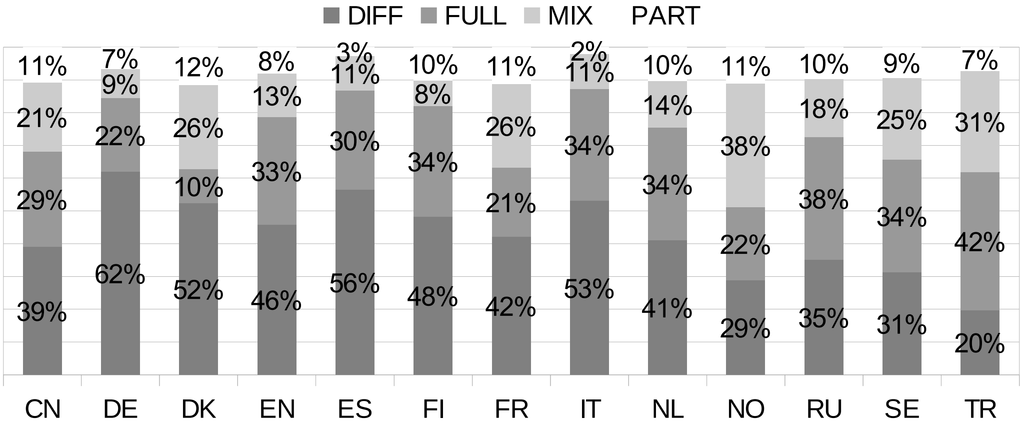

Figure 2 – Distribution of the correction response types

4.2 Normality Test

Shapiro Wilk normality test of the delta features of both datasets (VIC and IS09), independent of language, showed that they are not represented with a normal distribution, as is usually true for real data. In the case of VIC features, non-parametric one-sample Wilcoxon test applied to all the delta features, independent of language, showed that the minimum *f0* is significantly higher for the correction turns ($p<0.05$), higher maximal velocity ($p<0.001$), longer speech duration with unchanged phonation time ($p<0.001$), lower speech rate ($p<0.001$), lower articulation rate ($p<0.001$) and longer average syllable duration ($p<0.001$).

Similarly, for the IS09 delta features, it was observed that 202 features have non zero median values. Most notable, like in VIC features, are: the *f0* arithmetic mean (lower, $p<0.01$), as well as, the offset (lower, $p<0.001$) and the slope (greater, $p<0.001$) of its linear regression model. For RMS energy, the previously stated functionals are not significant predictors, but the maximum ($p<0.001$), minimum ($p<0.01$), the range ($p<0.001$) and the standard deviation ($p<0.001$) are significantly larger in the correction turns. The voicing probability also introduced changes in the slope (greater, $p<0.001$) and the offset (lower, $p<0.001$) of the contour, as well as, for ZCR mean (lower, $p<0.001$) and its standard deviation (lower, $p<0.001$).

Kruskal-Wallis rank sum test with language as a factor showed that, in the case of VIC features ($p<0.001$) except for final velocity and average syllable duration, and for IS09 except for 32 features, all other delta features belong to non-identical populations. These findings confirmed the presence of distinctive features mostly related to slower speech.

4.3 Linear Mixed Models

Better insight about the effects of the factors could be given by employing Linear Mixed Models [23] in R package for statistical computing [24]. The respective features were taken as dependent variables, the speaker as a random factor, the native language and the elicited corrections type as fixed factors, and all their possible interactions were included in the model.

The p-values for the fixed effects were calculated by the deviance table analysis (Type II Wald chi-square tests). For VIC features, the analysis with factor language revealed the significant predictors: maximum *f0*, mean *f0* with ($p<0.05$), minimum *f0*, final *f0* and maximum *f0* velocity ($p<0.01$), the excursion, mean intensity, speech and articulation rate and average syllable duration ($p<0.001$). For the correction type as a factor, all features except final *f0*, maximum *f0* velocity, and the excursion are significant ($p<0.001$). The statistical analysis further revealed significant interactions between language and the correction types ($p<0.001$) for all dependent variables except final velocity ($p>0.05$) and mean intensity ($p<0.05$).

Similarly, for the IS09 features the findings could be summarized as: for the language as a factor - 98 of the features, predominantly MFCC based functionals, are not significant, for the correction types, 88 of the features are not significant, notable among them *f0* range (excursion) and maximum *f0*. For the language and correction type interactions, 63 features, mostly the MFCC based features, are not significant predictors.

4.4 Summary of the Statistical Analysis

The Linear Mixed Model analysis showed that there is a significant influence of the fixed factors language and correction types, as well as their interaction at the delta features related to hyperarticulated speech. In general, the speakers raised their voice (pitch and intensity) 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).

5 Classification experiments and results

5.1 Experiments setup

The detection of different types of correction responses with present hyperarticulated speech is considered to be a multi-class classification problem. The datasets are characterized by a small number of unbalanced classes and a small number of observations per class. In all of the experiments, we are using the R package for statistical computing [24].

5.2 Classification methodology

The choice of an appropriate classification approach depends on a number of factors, 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 used SVM for classification, which seems well suited when applied on OpenSMILE derived acoustic-prosodic features.

Also, we assessed the usability of ANN in comparison to the other methods, considering the limitations of a rather small number of observations, inconsistent data set, and a quite large feature dimensions.

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 sets, as well as the validation folds. The experiments were repeated also in the case of 2 classes of corrections, deletions, and substitutions. The weighted guess classifier accuracy in the 3 class case is 0.455 and for the 2 class is 0.528.

5.2.1 Support Vector Machines

Although originally developed for binary classification, SVMs [25] are widely used also in multi-class recognition tasks. In order to achieve acceptable results, the correct choice of kernel parameters is very important. Before the results can be trusted, an extensive search has to be conducted on the hyper-parameter ranges to find the most optimal values (Table 2).

To train our SVMs, we took advantage of the R interface to the well-known LIBSVM library [26]. The Radial Kernel Function (RBF) was chosen because of the non-linearity nature of the classification problem and of its good general performance. The SVM was tuned over a range of the cost - C (10^{-4} to 10^{1}) and the gamma (10^{-9} to 10^{1}) parameters.

Table 2 – Results of classification experiments with the best performing SVM models

dataset	n	C	gamma	UAR	std	precision	std	recall	std	F1	std
VIC	3	1	1	0.605	0.024	0.394	0.010	0.437	0.089	0.561	0.025
	2	1	1	0.662	0.009	0.592	0.009	0.646	0.015	0.582	0.010
IS09	3	10	0.001	0.687	0.010	0.608	0.020	0.665	0.021	0.627	0.015
	2	10	0.001	0.731	0.015	0.720	0.012	0.717	0.013	0.718	0.012

5.2.2 Artificial Neural Networks

Different ANN models were trained through the same feature sets. The R interface to Keras [27], the neural network API was employed, with the Tensorflow [28], 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 (Table 3). The topology consisted of fully connected layers with an equal number of hidden units, leaky ReLU activation [29] function and, an L2 kernel regularizer. The output layer has softmax activation and two or 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 Adagrad [30] 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 a batch size of 256. The maximum number of epochs was set to 50 with at least 30 epochs as a condition for early stopping when there is no further improvement in the loss function of the validation set.

Table 3 – Results of classification experiments with the best performing ANN models

dataset	n	layers	units	UAR	std	precision	std	recall	std	F1	std
VIC	3	3	1100	0.549	0.005	0.551	0.018	0.491	0.007	0.493	0.012
	2	3	1100	0.655	0.015	0.654	0.013	0.647	0.013	0.645	0.014
IS09	3	1	2100	0.664	0.014	0.651	0.016	0.600	0.011	0.616	0.012
	2	1	1100	0.719	0.007	0.716	0.006	0.708	0.007	0.709	0.007

6 Conclusions

From the results of the statistical analysis, it could be clearly seen that there are distinctive acoustic-prosodic features associated with hyperarticulated speech in correction dialogue acts. For the classification experiments, we used SVM and ANN for multi-class classification of correction types. Many similar studies are dealing with classification of para-linguistic aspects in dialogue turns, most of them as binary classification tasks, except in the cases, where an adequate amount of data is available.

The achieved results were analyzed and compared in terms of unweighted accuracy, precision, recall, and F1 score. The best performing models, achieved better accuracy than the baseline weighted guess classifier. The ANN models did perform reasonably well despite the relatively small amount of observations and a larger number of features. When the task was reformulated as binary classification (deletions and substitutions errors) the ANN model provided results which are comparable with those obtained in similar tasks and on different speech databases.

7 References

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

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

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

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

[5] K. Yoshino, T. Hiraoka, G. Neubig, 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

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

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

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

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

[10] E. Shriberg, A. Stolcke, D. Jurafsky, N. Coccaro, M. Meteer, R. Bates, P. Taylor, K. Ries, R. Martin, C. Van Ess-Dykema: *Can Prosody Aid the Automatic*

Classification of Dialog Acts in Conversational Speech?, In: *Language and Speech*, 41:439-487, 1998

[11] 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.

[12] J. HIRSCHBERG, D. LITMAN, M. SWERTS: *Prosodic and other cues to speech recognition failures, In: Speech Communication*, 43(1-2):155–175, 2004.

[13] S. OVIATT, *Modeling hyperarticulate speech during human-computer error resolution*, In: *Proc. of the Int. Conference on Spoken Language Processing*, pp. 797-800, 1996

[14] M. SWERTS, D.J. LITMAN, J. HIRSCHBERG: *Corrections in spoken dialogue systems*, In: *INTERSPEECH*, pp. 615-618, 2000.

[15] A. FANDRIANTO, M. ESKENAZI, *Prosodic entrainment in an information-driven dialog system*, In: *13th Annual Conference of the Int. Speech Communication Association*, 2012.

[16] 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*, In: *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, New Orleans, LA, pp. 4985-4989.

[17] I. WENDLER, A. JATHO, I. KRALJEVSKI, M. WENZEL, *Nutzerzentrierter Entwurf von Multimodalen Bedien-konzepten*, In: *28. Konferenz Elektronische Sprach-signalverarbeitung 2017, Universität des Saarlandes, Saarbrücken*, 15.-17. März 2017.

[18] P. BOERSMA, *Praat, a system for doing phonetics by computer*, In: *Glot International* 5:9/10, 341-345, 2001.

[19] N.H. DE JONG, T. WEMPE, *Praat script to detect syllable nuclei and measure speech rate automatically*, In: *Behavior research methods*, 41 (2), 385-390, 2009.

[20] Y. XU, *ProsodyPro - A Tool for Large-scale Systematic Prosody Analysis*, In: *Proc. of Tools and Resources for the Analysis of Speech Prosody (TRASP 2013)*, Aix-en-Provence, France. 7-10, 2013.

[21] F. EYBEN, M. WÖLLMER, B. SCHULLER, *Opensmile: the Munich versatile and fast open-source audio feature extractor*, In: *Proc. of the 18th ACM international conference on Multimedia*. ACM, pp. 1459-1462, 2010.

[22] G. BELLER, N. OBIN, X. RODET, *Articulation degree as a prosodic dimension of expressive speech*, In: *4th International Conference on Speech Prosody*, 681-684, 2008.

[23] D. BATES, M. MAECHLER, B. BOLKER, S. WALKER, *Fitting Linear Mixed-Effects Models Using lme4*, In: *Journal of Statistical Software*, 67(1), 1-48, 2015

[24] 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.

[25] V. VAPNIK, *Statistical learning theory*, New York: Wiley, 1998

[26] C.C. CHANG, 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.

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

[28] M. ABADI, P. BARHAM, J. CHEN, Z. CHEN, A. DAVIS, J. DEAN, M. DEVIN, S. GHEMAWAT, G. IRVING, M. ISARD, M. KUDLUR, *TensorFlow: A System for Large-Scale Machine Learning*, In: *OSDI*, Vol. 16, pp. 265-283 (2016).

[29] A.L. MAAS, A.Y. HANNUN, A.Y. NG, *Rectifier nonlinearities improve neural network acoustic models*, In: *Proc. icml* (Vol. 30, No. 1, p. 3), June 2013

[30] J. DUCHI, E. HAZAN, Y. SINGER, *Adaptive subgradient methods for online learning and stochastic optimization*, In: *Journal of Machine Learning Research*, 12(Jul), pp.2121-2159, 2011