Transition-Relevance Places Machine Learning-Based Detection in Dialogue Interactions

Stanislav Ondas*, Matus Pleva, Silvia Bacikova
Department of Electronics and Multimedia Communications,
Faculty of Electrical Engineering and Informatics, Technical University of Kosice,
Letna 9, 04120 Kosice, Slovakia,
**stanislav.ondas@tuke.sk; matus.pleva@tuke.sk; silvia.bacikova@student.tuke.sk*

***Abstract*—A transition-relevance place (TRP) represents a place in a conversation where a change of speaker can occur. The appearance and use of these points in the dialogue ensures a correct and smooth alternation between the speakers. In the presented article, we focused on the study of prosodic speech parameters in the Slovak language, and we tried to experimentally verify the potential of these parameters to detect TRP. To study turn-taking issues in dyadic conversations, the Slovak dialogue corpus was collected and annotated. TRP places were identified by the human annotator in the manual labelling process. The data were then divided into chunks that reflect the length of the interpausal dialogue units and the prosodic features were computed. In the Matlab environment, we compared different types of classifiers based on machine learning in the role of an automatic TRP detector based on pitch and intensity parameters. The achieved results indicate that prosodic parameters can be useful in detecting TRP after splitting the dialogue into interpausal units. The designed approach can serve as a tool for automatic conversational analysis or can be used to label large databases for training predictive models, which can help machines to enhance human-machine spoken dialogue applications.**

***Index Terms*—Conversation analysis; Dialogue initiative; Transition-relevance places; Prosodic features; Classification.**

I. INTRODUCTION

Since it is difficult to speak and listen at the same time, dialogue interlocutors must somehow coordinate who is speaking and who is listening at any given moment. Turn-taking rules determine the order in which individual actions are to be performed and by whom. Even though human-machine conversational systems in various forms are becoming more and more common, the turn-taking in these systems is still not enough fluent and natural. These systems often tend to incorrectly interrupt the user or have a very long response delay. Thus, the modelling of turn-taking process during dialogue interactions is still largely a subject of active research [1].

Dialogue exchanges and taking turns between speakers imagines the basic building blocks of social interaction. Coordination of turn-taking is based on the personal analysis of the current speaker by the co-participant, i.e., the current listener, and on the identification of the possible termination of the speaker's speech (End of Turn (EOT)) based on syntactic, pragmatic, and prosodic cues. The timing of dialogue exchange in conversations can be considered extremely fast due to the cognitive demands of speakers to understand, plan, and execute a turn in real time [2].

A commonly used term in conversation analysis is the turn-constructional unit (TCU). It is the basic building block from which individual dialogue exchanges between speakers are constructed. An example of a TCU is a sentence, phrase, conjunction, or simple one-word answer. The end of any TCU represents the point at which the next speaker may express interest in taking the initiative in the dialogue. These endpoints are defined as transition-relevance places (TRPs), which means a possible transition place from one speaker to another. However, it should be noted that not every end of the TCU must have a transition. For this reason, these points are characterised as relevant, but not necessary, for the transition. The occurrence and use of TRP points in the dialogue ensures smooth switching of speakers. Thus, the TRP makes it easy for each participant to recognise when they will be able to start or end a particular TCU. For this reason, the TRP must be clearly predictable for each listener to achieve a smooth transition between speakers. The result of such a transition is the minimisation of gaps and overlaps between individual turns. A complicating factor in examining speaker transition points is that TRP cannot be directly observed in the data. We can only observe true turnovers, which could therefore be considered a subset of TRP. However, reversals can also occur where there is no TRP [3]. The relation between turn-constructional units and transition-relevance places is illustrated in Fig. 1.

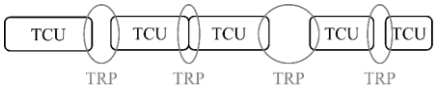

Fig. 1. Organisation of the TCU units and the TRP places.

An important concept in the analysis of the change of initiative is feedback (backchannel (BC)). Feedback can be characterised as a statement or speech by a participant in the role of a listener, but without the intention to take the initiative in speaking from the current speaker (without the intention to take over the role of the speaker). These feedback signals mostly show agreement, understanding, and also encourage the current speaker to continue

Manuscript received 21 January, 2023; accepted 29 March, 2023.
This research was supported by the SRDA under Grants No. APVV SK-TW-21-0002, APVV-15-0517; SGA MESRS under Grants No. VEGA 1/0753/20, VEGA 2/0165/21; and MESRS under Grant No. 2020-1-BG01-KA202-079200.

(continuers). Feedback usually occurs during short pauses in the speaker's speech and usually lasts less than a second. BC utterances primarily signal continued attention, approval, or various emotional responses. Feedback has a special status within the turn-taking rules because it occurs quite frequently during conversation, but is nevertheless not considered a suitable place for a change of speaker. Just as classical speaker turns occur after certain cues in speech (turn-yielding cues), the timing of feedback is related to certain cues that indicate it (backchannel inviting cues). We perceive the places that follow just after these stimuli to be relevant for the location of the feedback by the current listener and we call them a "place relevant for the backchannel" (backchannel-relevant place (BRP)) [1], [4].

To identify TCUs and TRPs as easily as possible, speech technology researchers have found it convenient to segment speech into interpausal units (IPUs). These are segments of spoken speech from one speaker without any silence exceeding a certain value (e.g., 200 ms). An utterance or turn is then typically defined as a sequence of IPUs from one speaker that are not interrupted by IPUs from another speaker. Silence between two IPUs from the same speaker is usually called a "pause", while silence between IPUs from different speakers is called a "gap" [1].

Several research studies have been conducted to examine different sets of features and models for predicting turn-taking, end-of-turn prediction, and TRP detection. Most of the described feature sets are based on prosodic features and energy features. The fundamental features are the fundamental frequency (*F0*) and power [5]–[8]. In addition to acoustic features, linguistic and multimodal features were used as a more complex representation of turn-management cues [9], [10].

Regarding prediction models, they were usually based mainly on conditional random fields [10], support vector machines [11], or neural networks [12]. Currently, the most preferred approaches are based on recurrent neural networks (long short-term memory (LSTM)). The advantage of the approaches mentioned above is that they consider long context of the input, and this way enables high accuracy to be achieved [9], [13]–[18]. In [19], the authors presented an approach based on the Izhikevich neuron model-based spiking neural network (SNN).

However, most of the mentioned papers focus on the end-of-turn prediction, or on TRP detection. These two tasks are closely related to each other and the same or similar feature sets and predictive approaches can be applied. In [17], the authors use the LSTM-based approach also to detect TRP to enhance the turn-taking prediction. In our experiments, we have tried to compare the simple machine learning-based classifiers, with a reduced set of acoustic features, with modern neural network-based approaches and their performance.

The paper is organised as follows. Section II describes the role of prosody in the turn-taking process and intensity and pitch as the fundamental prosodic features. Section III describes the approach, which we applied to create an automatic TRP detector, including a description of the dialogue corpus, its annotation, feature extraction, and the tools and algorithms used for training classifiers. In the next section, results of performed experiments are shown, followed by a brief discussion in the conclusion section.

II. SPEECH PROSODY IN TURN-TAKING

The role of prosody in turn-taking has become the subject of great interest and controversy. Prosody refers to the nonverbal aspects of speech, including intonation, loudness, rate of speech, and the like. It has been found to serve many important functions in conversation, including the evaluation of importance, syntactic ambiguity, attitudinal reactions, uncertainty, or shifts within topics.

Regarding intonation, studies in different languages have found that a stable level of intonation (in the middle of the speaker's fundamental frequency range) near the end of the IPU tends to serve as a cue for turn-holding. While rising or falling pitch can be found in contexts where there are indications of turn-yielding.

The intensity of the voice also carries a certain informational value for the detection of TRP locations. Speakers tend to lower their voice as they approach the limits of potential turnover, while speech has a higher intensity before pauses occur within the speech. Several studies have also investigated the role prosody plays in eliciting feedback and how these cues differ from cues for turn-taking. They found that feedback tends to arrive about 200 ms after the low-tone region. On the other hand, they also found that IPUs immediately before feedback showed a clear tendency towards final rising intonation, as well as higher intensity. These somewhat conflicting findings may be explained by language differences within the studies conducted. However, it is necessary to note that in no available study was data analysis carried out in Slovak.

Regardless of the role of prosody in turn-taking, prosody can provide important cues from the perspective of a conversational system. Since conversational systems do not have the same computational/cognitive limitations and do not need to pre-prepare a response to the extent that humans do, they could make greater use (detection and generation) of dialogue cues [1].

Pitch and intensity can be identified as fundamental prosody features. The pitch is defined by the fundamental frequency (*F0*) of the vibrations of the vocal cords. This frequency is specific to each speaker due to differences in the physical structure of the speaker's vocal cords and modulates the melody of the utterance. *F0* is also an important feature that characterises individual speakers, their gender, but also higher-level characteristics, e.g., their emotional state [20].

Speech intensity is generally recognised as one of the basic prosodic parameters. The term *intensity* is often replaced by the term *amplitude* or *loudness*. Intensity is the basic element of amplitude and is defined as the force transmitted by sound waves per unit area. Auditory perception of intensity is usually expressed in decibels [dB]. However, most linguists do not pay significant attention to intensity as a characteristic, but it plays a significant role in the definition of a syllable, which says that the syllable corresponds to the peak of intensity. Intensity is a demarcation function at different levels. Its value during a given time interval can be used to detect pauses, and thus can separate speech from nonspeech sequences. Also, the mutual relationship between intensity and fundamental

frequency can be considered as a certain physiological basis. Intensity and fundamental frequency are controlled by the same mechanisms, such as increases in pulmonary effort and subglottal pressure, vocal fold tension, etc. It can be assumed that a higher fundamental frequency is generally correlated with an increase in intensity. Similarly, the decrease in *F0* at the end of sentences is associated with a decrease in intensity [21].

In the proposed paper, we focused on the predictive force of these fundamental prosody features, pitch and intensity, to predict TRP place as a simple way to enhance turn-taking in case of human-machine interaction.

III. Automatic TRP Detection

The TRP places are often located in the pauses inside the turn-constructional units (TCU). The location of a possible TRP place can be predicted by the detection of turn-yielding cues - "events from acoustic, prosodic or syntactic sources, inter alia, produced by the speaker when approaching the potential end of a conversational turn, that may be used by the listener to detect, or even anticipate, an opportunity to take the floor" [22].

In our work, we focus only on prosodic cues, which are given by the actual speaker, toward the listener. Moreover, from the group of prosodic features, we selected pitch and intensity. These labelled data were used to train different machine learning classifiers using the Matlab Classification Learner tool and evaluated.

A. Dialogue Corpus and Annotation

The dialogue corpus, which we used to research turn-taking mechanisms, consists of audio/video recordings of dyadic human-human conversations in Slovak with a specific annotation. The corpus consists of recordings of the television discussion session "Pod lampou", where the discussion topics are predominantly focused on politics. The corpus contains eight television shows, with totally nine speakers (one moderator, eight different guests). Each recording consists of an audio and video file and with the annotation file.

Data were annotated using the annotation tools Transcriber [23], Anvil [24], and Elan [25]. Initially, transcriptions were created in the Transcriber tool. Then, due to our other research task, we processed the data in Anvil (Fig. 2) to add overlapping speech annotation (see [26]).

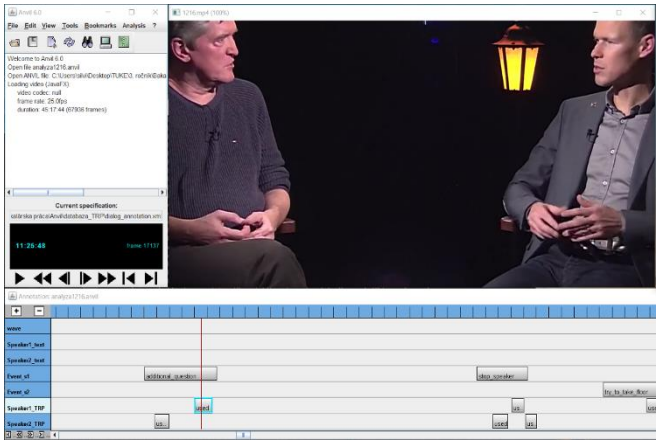

Fig. 2. An example of data and annotation in the Anvil tool.

Then, we switch our data to Elan tool to add annotation of TRP points. The reason of moving to Elan tool was that it enables us to localise TRP points with higher accuracy.

The TRP points were annotated by adding two different mark-ups into the data, which are placed into the point of possible TRP. We distinguished two situations. In the first, we identified TRP points that the listener used to take a turn (used TRP). The "unused TRP" mark-up was added in case of a possible TRP point identified by the annotator. In this case, the annotation is highly subjective, because there do not exist any described rules where the listener can take the floor.

The example of annotation in Elan tool can be seen in Fig. 3.

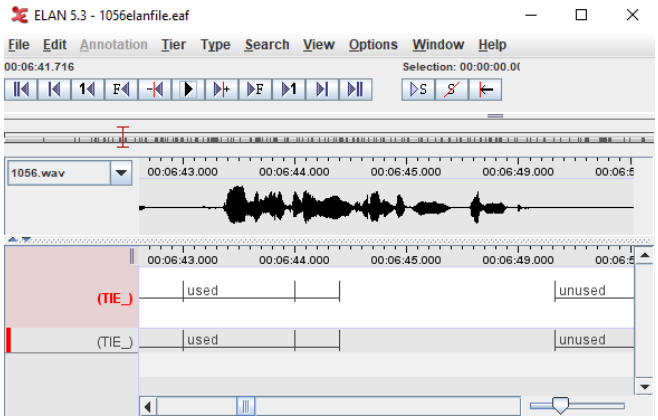

Fig. 3. An example of annotation data in Elan tool (without video player).

Using these rules, all recordings in the corpus were manually labelled. The occurrences of types of TRP in recordings are described in Table I.

TABLE I. TRP OCCURRENCE IN DATASET.

Recording	Duration	1. Interlocutor (moderator)		2. Interlocutor (guest)	
		Used TRP	Unused TRP	Used TRP	Unused TRP
1056	0:55:24	118	11	123	30
1058	1:17:15	54	5	62	51
1059	1:15:06	111	5	107	21
1062	1:04:05	92	6	91	26
1216	0:45:17	69	8	66	6
1217	0:42:32	33	1	32	14
1219	0:58:06	95	2	92	27
1305	0:54:54	82	1	81	18

B. Feature Extraction

In [7], average duration of IPU segments was measured in Slovak dialogue corpus, and the resulting value was approximately 1.32 seconds. In our experiments, we took this value as a basis of the analysed chunk, but according to observation of our data, we also took into each chunk 180 ms of pause segment. In this way, we expand the chunk duration to 1.5 s. By adding part of the pause between IPU segments, we tried to eliminate inaccuracies in the annotation of the data.

Each chunk of data was labelled with a *TRP*/*NO-TRP* annotation. Using the Praat tool [27], we gradually extracted prosodic features frame by frame.

The data chunks were parameterised according to the procedure described by in [28]. Each data chunk was split into 30 ms frames. For each frame, the values of intensity and fundamental frequency *F0* were computed. In the case

of *F0* values, in the case of unvoiced frames, a *0* value was inserted into the vector. The data vector, which represents each chunk of data, consists of 48 intensity values and 48 *F0* values.

In Fig. 4, the intensity and *F0* curves for the segment, with a possible TRP at its end, can be seen.

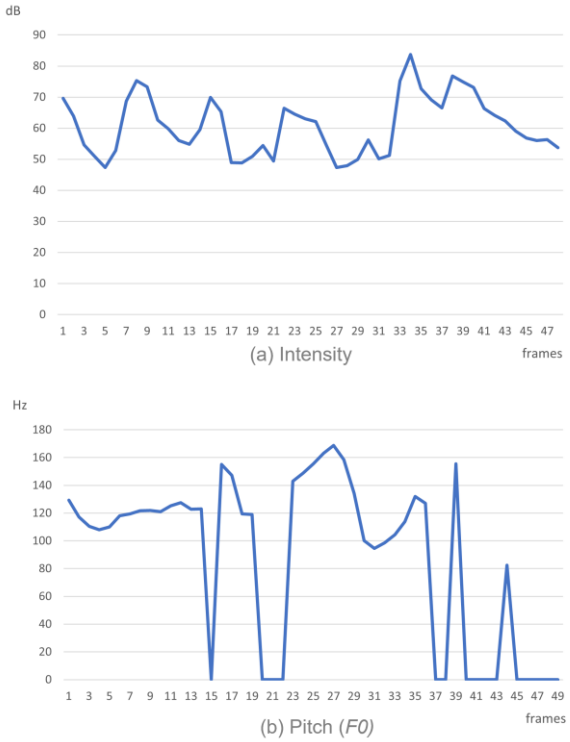

Fig. 4. (a) Intensity and (b) pitch of the IPU segment with subsequent TRP.

In Fig. 4, the intensity and *F0* curves for the segment, without possible subsequent TRP at its end, can be seen for comparison with the opposite situation depicted in Fig. 5.

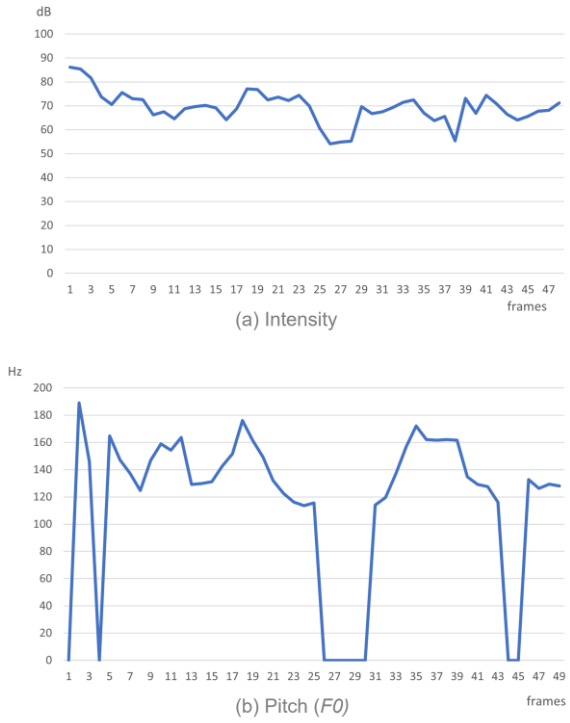

Fig. 5. (a) Intensity and (b) pitch of the IPU segment with NO subsequent TRP.

The data were then organised into a feature matrix and split into the training and testing part (75 %/25 %).

C. Algorithms and Tools

Supervised machine learning algorithms were selected for the automatic TRP detection task, where labelled data are used to train classifiers. The goal of classification is to learn a classification rule, based on which it would be possible to further implement automatic data determination with a certain accuracy.

The Matlab Classification Learner tool [29] was used to train and test classifiers. It allows to examine the selected data, select functions, specify validation schemes, training models, and evaluate the obtained results. Classification Learner performs automated training to find the best type of classification model including decision trees, discriminant analysis, support vector machines, logistic regression, nearest-neighbour classifiers, Naive Bayes classifiers, etc. Supervised machine learning makes it possible to perform automatic classification given a known set of input data (observations or examples) and known responses to the data (e.g., labels or classes). The given input data are used to train a model that generates predictions for the response on the new data. Furthermore, it allows the selected trained model to be exported to the Matlab workspace or to generate Matlab source code to recreate the trained model.

D. Training Classifiers

For the automatic classification of TRP places in spoken dialogue, data from three speakers was used to train classifiers. At the beginning, separate models were trained for each individual speaker. Later, we investigated the performance of the automatic classification also for the case where the input data represents a combination of data from these speakers. We also investigated a combination of training data from male and female interlocutors.

During training all available classifiers, the Classification Learner tool continuously evaluates the validation accuracy (Accuracy), based on which it is possible to get an immediate idea of the classification accuracy of each model. For each type of classifier used, the tool trains several types of the given classifier. For each classifier, we only considered classifiers with the best validation accuracy. The validation accuracy is calculated on a data set that is not used directly in training, but is used (during the training process) to pre-validate the performance of the model.

For each trained model, we evaluated the accuracy of the model using the "*Precision*", "*Recall*", and "*F-score*" formulas:

$$Precision = TP / (TP + FP), \tag{1}$$

where *TP* is *True Positives* and *FP* is *False Positives*,

$$Recall = TP / (TP + FN), \tag{2}$$

where *FN* is *False Negatives*, and

$$F - score = 2 \times \frac{Precision \times Recall}{Precision + Recall}, \tag{3}$$

which combines *Precision* and *Recall*.

These functions are used to evaluate the prediction accuracy of each model. In the field of pattern recognition, information retrieval, and classification (machine learning), precision and recall are performance metrics that apply to data obtained from a set of data samples. The *F-score* measures the accuracy of the classification, combining both precision and recall metrics.

IV. RESULTS

Totally 25 different classifiers were trained using Matlab Classification Learner with five different groups of data, as shown in Table II.

TABLE II. DATASETS FOR TRAINING.

	Train samples	**Test samples**
1 - Speaker (MALE)	128	42
2 - Speaker (MALE)	218	72
3 - Speaker (FEMALE)	352	118
1 + 2	346	114
1 + 2 + 3	698	232

The complete list of trained classifiers can be found in Table III.

TABLE III. CLASSIFIERS LIST.

Classifier	**Classifier**
Fine Tree	Coarse Gaussian SVM
Medium Tree	Fine KNN
Coarse Tree	Medium KNN
Linear Discriminant	Coarse KNN
Quadratic Discriminant	Cosine KNN
Logistic Regression	Cubic KNN
Gaussian Naive Bayes	Weighted KNN
Kernel Naive Bayes	Boosted Trees
Linear SVM	Bagged Trees
Quadratic SVM	Subspace Discriminant
Cubic SVM	Subspace KNN
Fine Gaussian SVM	RUSBoosted Trees
Medium Gaussian SVM	-

Table IV summarises the accuracy (ACC) results of the individual types of the group of 7 best classifiers for interlocutors and their combinations. For the case of the 1st speaker, the decision tree classifier and the ensemble classifier achieved the same best result of 98.4 %. For the case of the classification of the 2nd speaker, the Ensemble classifier achieved the best result of 92.7 %. For the case of the 3rd speaker, the Ensemble classifier achieved the highest accuracy of 94.0 %. In the case of the simultaneous classification of the 1st and 2nd speaker, the Ensemble classifier achieved the best result of 93.6 %. Finally, in the case of the simultaneous classification of the 1st, 2nd, and 3rd speakers, the best result of 94.4 % was also achieved by the same classifier.

TABLE IV. AUTOMATIC TRP CLASSIFIERS RESULTS - ACC.

Classifier	**Interlocutors**				
	1	**2**	**3**	**1 + 2**	**1 + 2 + 3**
Decision Tree	**98,4** %	86,2 %	92,9 %	93,4 %	92,3 %
Linear Discriminant	64,1 %	84,9 %	87,8 %	88,2 %	92,3 %
Logistic Regression	64,1 %	81,2 %	85,5 %	83,8 %	85,8 %
Naive Bayes	95,3 %	90,4 %	91,8 %	91,0 %	90,8 %
SVM	96,1 %	91,7 %	92,9 %	92,5 %	92,8 %
KNN	92,2 %	90,8 %	91,2 %	89,9 %	92,0 %
Ensemble	**98,4 %**	**92,7 %**	**94,0 %**	**93,6 %**	**94,4 %**

The Ensemble classifier uses the random forest ensemble method. This machine learning method for classification works by generating several decision trees at training time. For classification tasks, the output of the random forest is the class selected by most decision trees.

Classification Learner enables one to show the receiver operating characteristic (ROC) curve graph, which enables one to observe performance of the classification model at all thresholds of classification. This curve combines two parameters: the true positive rate (TPR) and the false positive rate (FPR).

Figure 6 shows the ROC curve for the best classifier trained on the data of all interlocutors. In this case, the TPR value is 0.95, while the FPR value is 0.06.

The area under the curve (AUC) value enables us to better evaluate the performance of the classifier. It represents the degree of data separability by a given classifier. The higher the AUC, the better the model will be able to distinguish between classes. Here, the AUC value for the best model is 0.98, which means that model is perfect in the classification task.

In the next step, we tested the classifier with the best result in the Matlab working environment. We again calculated the classification accuracy evaluation characteristics, which achieved the following results: *Precision* was equal to 0.9443, *Recall* was 0.9440, and the overall *F-score* was 0.9441.

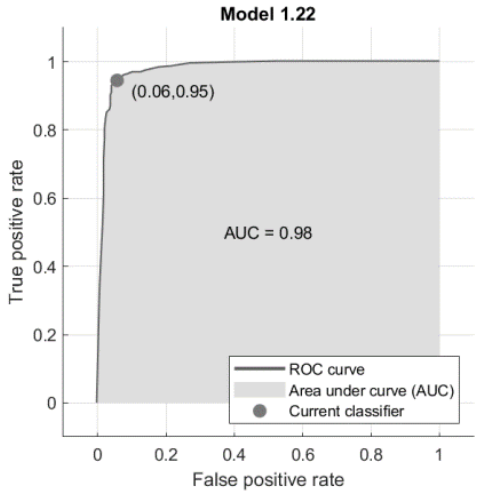

Fig. 6. ROC curve of the best classifier trained for all speakers.

Subsequently, we tested the classification model by applying it to a set of test data. After performing the automatic prediction, we observed specific cases of model inaccuracy. Of a total of 116 samples, including “yes_TRP” cases, the “A” labels were correctly assigned in 111 cases. In 5 samples, the model predicted the wrong label “N”. The second group consisted of samples representing “non_TRP” cases, where 109 labels of “N” were correctly assigned out of a total of 116 samples. It follows that in 7 cases, a wrong prediction of TRP labels was made by assigning the label “A”. From the evaluated results, we can conclude that the inaccuracies created during the implementation of automatic prediction correspond to the level of validation accuracy (ACC) of the classifier of 94.4 %.

Based on the results obtained for individual classifiers, we can conclude that, for input data from three speakers at the

same time, the Ensemble classifier achieves the best results.

V. CONCLUSIONS

The proposed paper investigates the role and effectiveness of using simple prosodic features, fundamental frequency of speech and intensity to classify interpausal units (IPU) into two categories, those that may be followed by a transition-relevance place (TRP) and those that are not likely to be.

In addition, 25 different classifiers, from the category of supervised machine learning methods, were trained and evaluated for this classification task, which resulted in the selection of the group of classifiers with the best results. In most cases, the best results were achieved by the Ensemble classifier (approximately 94 %). Ensemble methods use multiple learning algorithms to obtain better predictive performance than could be obtained from any of the constituent learning algorithms alone [30]–[32]. Our results confirmed this definition.

All these results were obtained on the Slovak dialogue corpus, which contains eight recordings of television discussions, with overall duration of about 8 hours. We described the way, how to annotate such data (TRP points) and the approach, how to split data into chunks. According to the work of prof. Beňuš in [7], we defined the length of the chunk to be one and half seconds.

In [17], the result achieved by the authors on the TRP detection task was between 81.7 % and 91 % (ACC). Their TRP detector was based on a hierarchical LSTM model, which took prosodic and linguistic parameters as its input. Our significantly simpler detector with only prosodic feature set gave an accuracy of 94 %. Compared to the results achieved by the authors in [18] on a similar task of predicting turn-holding behaviour after a pause, our detector achieved better results (*F-score* 0.9441 vs. 0.825). Here, the authors relied on several prosody parameters, including the spectral flux. In this comparison, we were able to achieve similar results with significantly simpler feature set and classifier.

Although the achieved results show high accuracy, we are aware that the accuracy achieved is higher than what a real classification system would achieve. We have two interpretations of this claim. The first is that we trained our system only with the chunks followed by the pause segments. The reason for selecting only these segments was that the number of other segments (without succeeding pause) was significantly higher than the chunks with TRP points and the chunks without TRP point, and this fact could cause overtraining of the classifier. Therefore, we decided to focus only on chunks at the end of IPU segments, just before the occurrence of pause.

The other uncertainty about the results can be seen in the fact that perception of the possible TRP and of the turn-holding/turn-yielding cues is highly individual and subjective.

Despite described issues, we hope that the performed experiments can help in the selection of appropriate features and classifiers for the automatic TRP detection/prediction for the human-machine dialogue systems in many applications.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest.

REFERENCES

[1] G. Skantze, "Turn-taking in conversational systems and human-robot interaction: A review", *Computer Speech & Language*, vol. 67, art. 101178, 2021. DOI: 10.1016/j.csl.2020.101178.

[2] J. Holler, K. H. Kendrick, M. Casillas, and S. C. Levinson, "Editorial: Turn-taking in human communicative interaction", *Frontiers in Psychology*, vol. 6, p. 1919, 2015. DOI: 10.3389/fpsyg.2015.01919.

[3] S. E. Clayman, "Turn-constructional units and the transition-relevance place", in *The Handbook of Conversation Analysis*. Oxford, U.K., Wiley-Blackwell, 2012, pp. 150–166. DOI: 10.1002/9781118325001.ch8.

[4] M. Kuswandi and Y. Apsari, "An analysis of pauses, overlaps and backchannels in conversation in vlog by Nessie Judge", *Project (Professional Journal of English Education)*, vol. 2, no. 3, pp. 282–291, 2019. DOI: 10.22460/project.v2i3.p282-291.

[5] O. Niebuhr, K. Gors, and E. Graupe, "Speech reduction, intensity, and F0 shape are cues to turn-taking", in *Proc. of the SIGDIAL 2013 Conference*, 2013, pp. 261–269. [Online]. Available: https://aclanthology.org/W13-4040/

[6] M. Zellers, "Pitch and lengthening as cues to turn transition in Swedish", in *Proc. of 14th Annual Conference of the International Speech Communication Association (INTERSPEECH 2013*), 2013, pp. 248–252. DOI: 10.21437/Interspeech.2013-77.

[7] A. Gravano, P. Brusco, and S. Benus, "Who do you think will speak next? Perception of turn-taking cues in Slovak and Argentine Spanish", in *Proc. of 17th Annual Conference of the International Speech Communication Association (INTERSPEECH 2016)*, 2016, pp. 1265–1269. DOI: 10.21437/Interspeech.2016-585.

[8] P. Brusco, J. M. Perez, and A. Gravano, "Cross-linguistic study of the production of turn-taking cues in American English and Argentine Spanish", in *Proc. of 18th Annual Conference of the International Speech Communication Association (INTERSPEECH 2017)*, 2017, pp. 2351–2355. DOI: 10.21437/Interspeech.2017-124.

[9] A. Maier, J. Hough, and D. Schlangen, "Towards deep end-of-turn prediction for situated spoken dialogue systems", in *Proc. of 18th Annual Conference of the International Speech Communication Association (INTERSPEECH 2017)*, 2017, pp. 1676–1680. DOI: 10.21437/Interspeech.2017-1593.

[10] M. Roddy, G. Skantze, and N. Harte, "Multimodal continuous turn-taking prediction using multiscale RNNs", in *Proc. of the 20th ACM International Conference on Multimodal Interaction (ICMI'18)*, 2018, pp. 186–190. DOI: 10.1145/3242969.3242997.

[11] J. Kane, I. Yanushevskaya, C. de looze, B. Vaughan, and A. Ni Chasaide, "Analysing the prosodic characteristics of speech-chunks preceding silences in task-based interactions", in *Proc. of 15th Annual Conference of the International Speech Communication Association (INTERSPEECH 2014)*, 2014, pp. 333–337. DOI: 10.21437/Interspeech.2014-79.

[12] N. G. Ward, O. Fuentes, and A. Vega, "Dialog prediction for a general model of turn-taking", in *Proc. of 11th Annual Conference of the International Speech Communication Association (INTERSPEECH 2010)*, 2010, pp. 2662–2665. DOI: 10.21437/Interspeech.2010-706.

[13] M. Roddy, G. Skantze, and N. Harte, "Investigating speech features for continuous turn-taking prediction using LSTMs", in *Proc. of 19th Annual Conference of the International Speech Communication Association (INTERSPEECH 2018)*, 2018, pp. 586–590. DOI: 10.21437/Interspeech.2018-2124.

[14] R. Masumura, T. Tanaka, A. Ando, R. Ishii, R. Higashinaka, and Y. Aono, "Neural dialogue context online end-of-turn detection", in *Proc. of the 19th Annual SIGdial Meeting on Discourse and Dialogue*, 2018, pp. 224–228. DOI: 10.18653/v1/W18-5024.

[15] D. Lala, K. Inoue, and T. Kawahara, "Evaluation of real-time deep learning turn-taking models for multiple dialogue scenarios", in *Proc. of the 20th ACM International Conference on Multimodal Interaction (ICMI'18)*, 2018, pp. 78–86. DOI: 10.1145/3242969.3242994.

[16] K. Hara, K. Inoue, K. Takanashi, and T. Kawahara, "Prediction of turn-taking using multitask learning with prediction of backchannels and fillers", in *Proc. of Interspeech 2018*, 2018, pp. 991–995. DOI: 10.21437/Interspeech.2018-1442.

[17] K. Hara, K. Inoue, K. Takanashi, and T. Kawahara, „Turn-taking prediction based on detection of transition relevance place", in *Proc. of 20th Annual Conference of the International Speech*

Communication Association (INTERSPEECH 2019), 2019, pp. 4170–4174. DOI: 10.21437/Interspeech.2019-1537.

[18] N. G. Ward, D. Aguirre, G. Cervantes, and O. Fuentes, "Turn-taking predictions across languages and genres using an LSTM recurrent neural network", in *Proc. of 2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 831–837. DOI: 10.1109/SLT.2018.8639673.

[19] S. Feng, W. Xu, B. Yao, Z. Liu, and Z. Ji, "Early prediction of turn-taking based on spiking neuron network to facilitate human-robot collaborative assembly", in *Proc. of 2022 IEEE 18th International Conference on Automation Science and Engineering (CASE)*, 2022, pp. 123–129. DOI: 10.1109/CASE49997.2022.9926441.

[20] M. Sigmund, "Statistical analysis of fundamental frequency based features in speech under stress", *Information Technology and Control*, vol. 42, no. 3, pp. 286–291, 2013. DOI: 10.5755/j01.itc.42.3.3895.

[21] E. Koffi, "A comprehensive review of intensity and its linguistic applications", *Linguistic Portfolios*, vol. 9, art. 2, 2020. [Online]. Available: https://repository.stcloudstate.edu/stcloud_ling/vol9/iss1/2

[22] A. Gravano and J. Hirschberg, "Turn-taking cues in task-oriented dialogue", *Computer Speech & Language*, vol. 25, no. 3, pp. 601–634, 2011. DOI: 10.1016/j.csl.2010.10.003.

[23] C. Barras, E. Geoffrois, Z. Wu, and M. Liberman, "Transcriber: Development and use of a tool for assisting speech corpora production", *Speech Communication*, vol. 33, nos. 1–2, 2001. DOI: 10.1016/S0167-6393(00)00067-4.

[24] M. Kipp, "Anvil - A generic annotation tool for multimodal dialogue", in *Proc. of the 7th European Conference on Speech Communication and Technology (Eurospeech 2001)*, 2001, pp. 1367–1370. DOI: 10.21437/Eurospeech.2001-354.

[25] P. Wittenburg, H. Brugman, A. Russel, A. Klassmann, and H. Sloetjes, "ELAN: A professional framework for multimodality research", in *Proc. of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, 2006, pp. 1556–1559. [Online]. Available: http://www.lrec-conf.org/proceedings/lrec2006/summaries/153.html

[26] S. Ondáš, M. Pleva, and J. Juhár, "Overlapping speech analysis in Slovak conversational interactions", in *Proc. of 2022 IEEE 20th Jubilee World Symposium on Applied Machine Intelligence and Informatics (SAMI)*, 2022, pp. 000401–000406. DOI: 10.1109/SAMI54271.2022.9780684.

[27] P. Boersma and V. van Heuven, "Speak and unSpeak with PRAAT", *Glot International*, vol. 5 no. 9/10, pp. 341–347, 2001. [Online]. Available: www.fon.hum.uva.nl/paul/papers/speakUnspeakPraat_glot2001.pdf

[28] M. Paľová and E. Kiktová, "Prosodic anticipatory clues and reference activation in simultaneous interpretation", *XLinguae: European Scientific Language Journal*, vol. 12, no. 1XL, pp. 13–22, 2019. DOI: 10.18355/XL.2019.12.01XL.02.

[29] "Classification Learner", The MathWorks, Inc. [Online]. Available: https://www.mathworks.com/help/stats/classificationlearner-app.html

[30] D. Opitz and R. Maclin, "Popular ensemble methods: An empirical study", *Journal of Artificial Intelligence Research*, vol. 11, no. 1, pp. 169–198, 1999. DOI: 10.1613/jair.614.

[31] R. Polikar, "Ensemble based systems in decision making", *IEEE Circuits and Systems Magazine*, vol. 6, no. 3, pp. 21–45, 2006. DOI: 10.1109/MCAS.2006.1688199.

[32] L. Rokach, "Ensemble-based classifiers", *Artificial Intelligence Review*, vol. 33, pp. 1–39, 2010. DOI: 10.1007/s10462-009-9124-7.