Recognising Perceived Task Difficulty from Speech and Pause Histograms

Ruth Janning, Carlotta Schatten, and Lars Schmidt-Thieme

Information Systems and Machine Learning Lab (ISMLL), University of Hildesheim
{janning,schatten,schmidt-thieme}@ismll.uni-hildesheim.de

Abstract. Currently, a lot of research in the field of intelligent tutoring systems is concerned with recognising student's emotions and affects. The recognition is done by extracting features from information sources like speech, typing and mouse clicking behaviour or physiological sensors. In former work we proposed some low-level speech features for perceived task difficulty recognition in intelligent tutoring systems. However, by extracting these features some information hidden in the speech input is loosed. Hence, in this paper we propose and investigate speech and pause histograms as features, which preserve some of the loosed information. The approach of using speech and pause histograms for perceived task difficulty recognition is evaluated by experiments on data collected in a study with German students solving mathematical tasks.

Keywords: Intelligent tutoring systems, perceived task difficulty recognition, low-level speech features, speech and pause histograms

1 Introduction

Automatic cognition, affect and emotion recognition is a relatively young and very important research field in the area of adaptive intelligent tutoring systems. Some research has been done to identify useful information sources and appropriate features able to describe student's cognition, emotions and affects. Those information sources can be speech input, written input, typing and mouse clicking behaviour or input from physiological sensors. In former work ([5], [6], [7]) we proposed low-level speech features for perceived task difficulty recognition in intelligent tutoring systems. These features are extracted from the amplitudes of speech input of students interacting with the system and contain for instance the maximal and average length of speech phases and pauses. However, by extracting those features some more fine granulated information contained within the sequence of speech and pause segments is loosed and the question arises if there is a way to create features which preserve the loosed information. Histograms contain much more information than only the maximal, minimal and average value. Hence, in this work we propose and investigate speech and pause histograms as features for perceived task difficulty recognition, i.e. for recognising if a student feels *over-challenged* or *appropriately challenged* by a task. Speech and pause histograms share the advantages of low-level speech features

(they do not inherit the error from speech recognition and there is no need that students use words related to emotions or affects, see also sec. 2) and avoid to lose information hidden in the sequences of speech and pause segments.

2 Related Work

For the purpose to recognise emotion or affect in speech one can distinct linguistics features, like n-grams and bag-of-words, and low-level features like prosodic features, disfluencies, e.g. speech pauses ([5], [6]), (see e.g. [17]) or articulation features ([7]). If linguistics features are not extracted from written but from spoken input, a transcription or speech recognition process has to be applied to the speech input before emotion or affect recognition can be conducted. Linguistic features for affect and emotion recognition from conversational cues were presented and investigated e.g. in [10] and [11]. Low-level features are used in the literature for instance for expert identification, as in [18], [13] and [8], for emotion and affect recognition as in [12] and [5], [6], [7] or for humour recognition as in [15]. The advantage of using low-level features like disfluencies is that instead of a full transcription or speech recognition approach only for instance a pause identification has to be applied before computing the features. That means that one does not inherit the error of the full speech recognition approach. Furthermore, these features are independent from the need that students use words related to emotions or affects. Another kind of features which is independent from the need that students use words related to emotions or affects are features gained from information about the actions of the students interacting with the system (see e.g. [9]) like features extracted from a log-file (see e.g. [2], [16], [14]). In [9] such kind of features is used to predict whether a student can answer correctly questions in an intelligent learning environment without requesting help and whether a student's interaction is beneficial in terms of learning. Also the keystroke dynamics features used in [4] belong to this kind of features. In [4] emotional states were identified by analysing the rhythm of the typing patterns of persons on a keyboard. A further possibility of gaining features is using the information from physiological sensors as for instance in [1]. However, bringing sensors into classrooms is time consuming and expensive and one has to cope with students' acceptance of the sensors.

3 Speech and Pause Histograms

As mentioned above, in this paper we investigate the ability of speech and pause histograms for perceived task difficulty recognition. How these speech and pause histograms are created from students' speech input is described in sec. 3.2 and the data which we used for our experiments is described in the next section.

3.1 Data

We conducted a study in which the speech and actions of ten 10 to 12 years old German students were recorded and their perceived task-difficulties were

Fig. 1. Graphic of the decibel scale of an example sound file of a student. The two straight horizontal lines indicate the threshold.

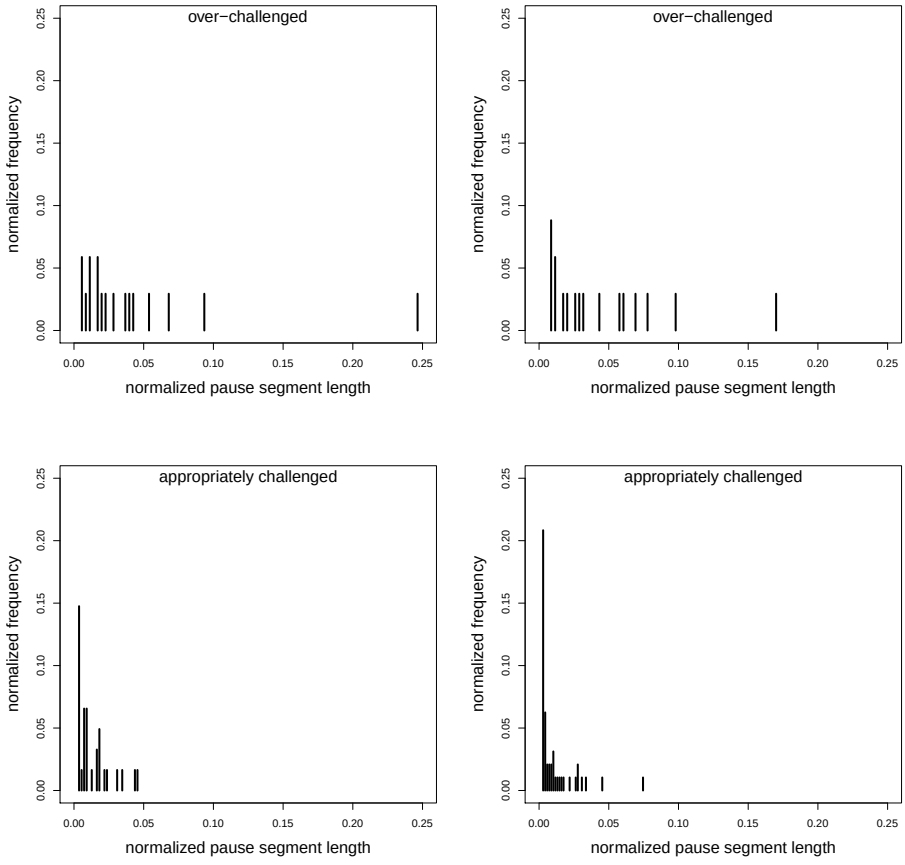

Fig. 2. Normalised pause histograms for a task of four different students, where two are labelled as *over-challenged* and the other two as *appropriately challenged*.

reported per task. The labelling of these data was done on the one hand concurrently by a human tutor and on the other hand retrospectively by a second reviewer (with a Cohen's kappa for inter-rater reliability of 0.747, $p < 0.001$). Divergences in the both labellings were clarified later on by discussions between the reviewers. During the study a paper sheet with fraction tasks was shown to the students and they were asked to paint – by means of a software for painting with a computer – their solution and they were prompt to explain aloud their observations and answers. The fraction tasks were subdivided into similar subtasks and covered exercises like assigning fractions to coloured parts of a circle or rectangle, reducing, adding or subtracting fractions and fraction equivalence. Originally, there were 10 tasks with 1 up to 10 subtasks but not each task was seen by each student. We made a screen recording to record the painting of the students and an acoustic recording to record the speech of the students. The screen recordings were used for the retrospective annotation. The acoustic speech recordings, consisting of 10 wav files with a length from 15 up to 20 minutes, were used to gain the speech and pause histograms. The data collection resulted in 36 examples (tasks) labelled with *over-challenged* (12 examples) or *appropriately challenged* (24 examples), respectively 48 examples (24 of class *appropriately challenged*, 24 of class *over-challenged*) after applying oversampling to the smaller set of examples of class *over-challenged* to eliminate the unbalance in the data.

3.2 Histograms for Classification

In the above mentioned study we observed that the children often exhibited longer pauses of silence while thinking about the problem when they were *over-challenged* or produced fewer and shorter pauses while communicating when they were *appropriately challenged*. Hence, in this paper we investigate information about pauses and speech segments within the speech input of students in connection with the perceived task difficulty. The first step to gain this information is to segment the acoustic speech recordings for identifying segments containing speech and segments corresponding to pauses. The most easy way to do this is to define a threshold on the decibel scale as done e.g. in [8]. For our study of the data we also used a threshold, which was estimated manually. The manual threshold estimation was done by extracting the amplitudes of the sound files, computing the decibel values and generating a graphic of it like the one in fig. 1. Subsequently, it was investigated which decibel values belong to speech and which ones to pauses to create from this information an appropriate threshold. By means of this threshold the pause and speech segments can be extracted. From the pause segments the pause histogram is generated by counting how often each possible pause length occur. This pause histogram is then normalised, to make the pause histograms of different speech inputs (of different students, different tasks and different lengths) comparable. The normalisation is done by dividing each occurring pause length by the length of the whole speech input as well as dividing the frequency of each occurring pause length by the number of all speech and pause segments, so that the resulting values stem

from the interval between 0 and 1. The same is done with the speech segments for generating the speech histogram. Examples of normalised pause histograms and speech histograms are shown in fig. 2 and fig. 3. The examples stem from the speech input for a task of four different students, where two were labelled as *over-challenged* and the other two as *appropriately challenged*. One can see some

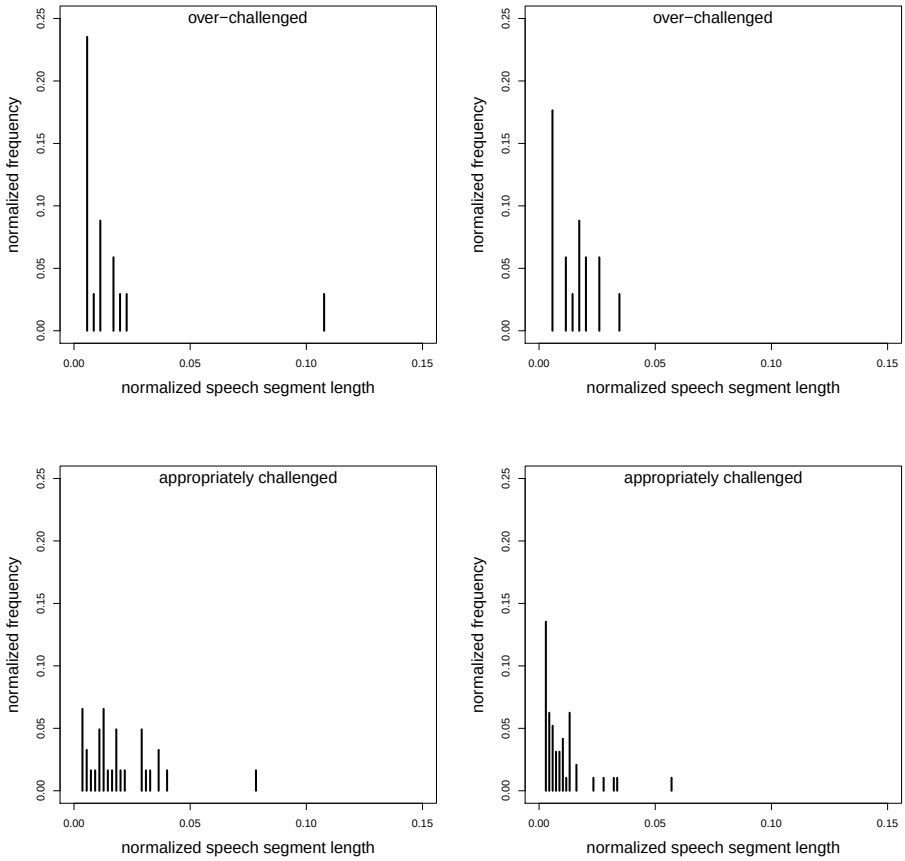

Fig. 3. Normalised speech histograms for a task of four different students, where two are labelled as *over-challenged* and the other two as *appropriately challenged*.

differences between the histograms of the *over-challenged* students and the *appropriately challenged* students as well as some similarities of the examples with the same label. The pause histograms of the *appropriately challenged* students show that there are a lot of very small pauses within their speech, but no very large pauses. The pause histograms of the *over-challenged* students in contrast

report long pauses and less smaller pauses than for the *appropriately challenged* students. In the speech histograms one can see that the *over-challenged* students used a lot of very small speech segments of the same length whereas for *appropriately challenged* students there is a large variance in the speech segment length. In the following section we investigate how these histograms can be used for classifying the speech input of a student for a task as either *over-challenged* or *appropriately challenged*.

4 Experiments

To investigate if the above described speech and pause histograms are applicable for distinguishing *over-challenged* and *appropriately challenged* students we conducted experiments with the perprocessing and settings described in the following section. The experimental results are reported in sec. 4.2.

4.1 Preprocessing and Experimental Settings

To be computationally comparable the normalised histograms still need to be preprocessed, or more explicitly generalised, as the set of possibly occurring segment lengths is infinite (it is a real value between 0 and 1). Hence, we divide the x-axis (the different normalised lengths of pause or speech segments) into a number of equal sized intervals, the *buckets*. Each occurring normalised segment length is then put into the bucket to whose interval it belongs. The number of buckets, or the bucket size respectively, is a hyper parameter and in the experiments we investigated different values for that parameter, i.e. we conducted experiments with 2 up to $1,000,000$ buckets (bucket size 0.5 up to 1.0E-6) where the numbers of buckets are multiples of the numbers by which 100 is divisible without remainder. A comparison of two different histograms can now be done by comparing the content of each bucket in both histograms, that means that for each bucket the normalised frequencies of segments belonging to that bucket are compared. In our experiments we compute the difference between two histograms by computing the differences between the frequencies in all buckets by means of the root mean square error (RMSE):

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{b}(b_i(H_x) - b_i(H_y))^2}{b}}\,, \tag{1}$$

where H_x and H_y are the two histograms to compare, $b_i(H_x)$ and $b_i(H_y)$ are the normalised frequency values belonging to bucket b_i of H_x and H_y and b is the number of buckets. For deciding to which class (*over-challenged* or *appropriately challenged*) a histogram belongs we applied the K-Nearest-Neighbour (KNN) approach. KNN (see e.g. [3]) classifies an example by a majority vote of its neighbours, that is the example is assigned to the class most common among its K nearest neighbours. These K nearest neighbours are the K closest training examples in the feature space. The *closeness* in our case is measured by means

of the RMSE. That is a histogram is assigned to that class to which the majority of the K closest (in terms of RMSE) histograms belongs. K is a further hyper parameter and also for that parameter we tried out different values, i.e. we conducted experiments with a number of 1 up to 35 neighbours where that value is an odd number less than the number of unique examples. For the evaluation we used a Leave-one-out cross-validation in the experiments. The results of our experiments with pause and speech histograms are discussed in the next section.

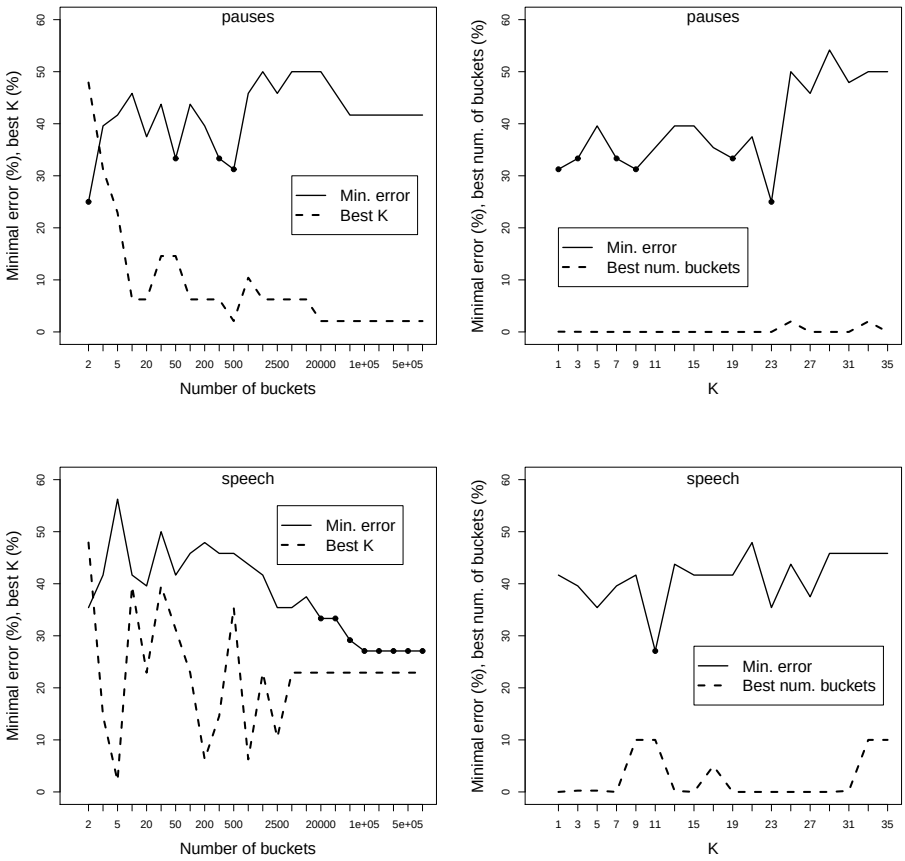

Fig. 4. Different numbers of buckets and different numbers K of neighbours mapped to the minimal classification error (%) and the belonging best value for K (% of the number of examples) and for the number of buckets (% of the max. number of buckets) for pause and speech histograms.

4.2 Experiments with Speech and Pause Histograms

As mentioned above, we conducted experiments with different numbers of buckets and different values for the K nearest neighbours. In fig. 4 we report the minimal classification error and the belonging best value of K for each bucket number as well as the the minimal classification error and the belonging best number of buckets for each value of K for the pause and the speech histograms. The *classification error* is the number of incorrectly classified histograms divided by the number of all histograms. The black dots in fig. 4 indicate the best results which are also reported in tab. 1 and 2. As one can see in fig. 4 for the

Table 1. Number of buckets, bucket size, K, classification error and F-measures of class *over-challenged* & *appropriately challenged* of the experiments with pause histograms with best result (classification error $< 34\%$, black dots in fig. 4).

Number of buckets	2	2	2	50	250	500
Bucket size	0.5	0.5	0.5	0.02	0.004	0.002
K	9	19	23	7	3	1
Error (%)	31.25	33.33	25.00	33.33	33.33	31.25
F-measure	0.57, 0.82	0.55, 0.80	0.67, 0.83	0.59, 0.63	0.59, 0.57	0.60, 0.71

Table 2. Number of buckets, bucket size, K, classification error and F-measures of class *over-challenged* & *appropriately challenged* of the experiments with speech histograms with best result (classification error $< 34\%$, black dots in fig. 4).

Number of buckets	20000	25000	50000	100000	200000	250000	500000	1000000
Bucket size	5.0E-5	4.0E-5	2.0E-5	1.0E-5	5.0E-6	4.0E-6	2.0E-6	1.0E-6
K	11	11	11	11	11	11	11	11
Error (%)	33.33	33.33	29.17	27.08	27.08	27.08	27.08	27.08
F-measure	0.57, 0.73	0.57, 0.73	0.62, 0.78	0.64, 0.77	0.64, 0.77	0.64, 0.77	0.64, 0.77	0.64, 0.77

pause histograms a smaller number of buckets delivers the best results whereas for the speech histograms the number of buckets has to be large, i.e. a more fine granulated division of the x-axis is needed for good results. The reason might be that the pause histograms of *over-challenged* and *appropriately challenged* students are easier distinguishable as in the pause histogram of an *over-challenged* student there are typically long pause segments which usually do not occur in the speech of *appropriately challenged* students (see also fig. 2). As fig. 3 shows, speech histograms of *over-challenged* and *appropriately challenged* students are not so easy to distinct. Tab. 1 and 2 show the results of the best choices for hyper parameter K and number of buckets and reports the classification error as well as the F-measures of both classes (*over-challenged* and *appropriately challenged*).

The F-measure is a value between 0 and 1 and the closer it is to 1 the better. It is the harmonic mean between the ratio of examples of a class c which are correctly recognised as members of that class (*recall*) and the ratio of examples classified as belonging to class c which actually belong to class c (*precision*). In our experiments the F-measures of class *appropriately challenged* are better than those of class *over-challenged*. The reason could be that originally there were more examples of class *appropriately challenged* and we just oversampled class *over-challenged* to receive a balanced example set. Nevertheless, the best classification errors of 25% and 27.08% and F-measures $0.67, 0.83$ and $0.64, 0.77$ in tab. 1 and 2 indicate that speech and pause histograms are applicable for perceived task difficulty recognition.

5 Conclusions and Future Work

We proposed and investigated speech and pause histograms, build from the sequences of speech and pause segments within the speech input of students, as features for perceived task difficulty recognition. To evaluate the approach of using the histograms for distinguishing *over-challenged* and *appropriately challenged* students we applied a K-Nearest-Neighbour classification delivering a classification error of 25% for pause histograms and 27.08% for speech histograms. Next steps will be to try out other classification approaches, for instance from time series classification. Furthermore, the information from the speech histograms and pause histograms could be combined to reach a better classification performance, e.g. by ensemble methods.

Acknowledgements. This work is co-funded by the EU project iTalk2Learn (www.italk2learn.eu) under grant agreement no. 318051.

References

1. Arroyo, I., Woolf, B.P., Burelson, W., Muldner, K. , Rai, D. and Tai, M.: A Multimedia Adaptive Tutoring System for Mathematics that Addresses Cognition, Metacognition and Affect. In International Journal of Artificial Intelligence in Education, Springer, Vol. 24, pp. 387–426 (2014)
2. Baker, R.S.J.D., Gowda, S., Wixon, M., Kalka, J., Wagner, A., Salvi, A., Aleven, V., Kusbit, G., Ocumpaugh, J. and Rossi,L.: Towards Sensor-Free Affect Detection in Cognitive Tutor Algebra. In Proceedings of the 5th International Conference on Educational Data Mining (EDM 2012), pp. 126–133 (2012)
3. Cover, T. and Hart, P.: Nearest neighbor pattern classification. EEE Transactions on Information Theory, Vol. 13(1), pp. 21–27, doi:10.1109/TIT.1967.1053964 (1967)
4. Epp, C., Lippold, M. and Mandryk, R.L.: Identifying Emotional States Using Keystroke Dynamics. In Proceedings of the 2011 Annual Conference on Human Factors in Computing Systems (CHI 2011), pp. 715–724 (2011)
5. Janning, R., Schatten, C., Schmidt-Thieme, L.: Multimodal Affect Recognition for Adaptive Intelligent Tutoring Systems. In Extended Proceedings of the 7th International Conference on Educational Data Mining (EDM 2014), pp. 171–178 (2014)

6. Janning, R., Schatten, C., Schmidt-Thieme, L.: Feature Analysis for Affect Recognition Supporting Task Sequencing in Adaptive Intelligent Tutoring Systems. In Proceedings of the European Conference on Technology Enhanced Learning (EC-TEL 2014), pp. 179–192 (2014)
7. Janning, R., Schatten, C., Schmidt-Thieme, L. and Backfried, G.: An SVM Plait for Improving Affect Recognition in Intelligent Tutoring Systems. In Proceedings of the IEEE International Conference on Tools with Artificial Intelligence (ICTAI) (2014)
8. Luz, S.: Automatic Identification of Experts and Performance Prediction in the Multimodal Math Data Corpus through Analysis of Speech Interaction. Second International Workshop on Multimodal Learning Analytics, Sydney Australia (2013)
9. Mavrikis, M.: Data-driven modelling of students interactions in an ILE. In Proceedings of the International Conference on Educational Data Mining (EDM 2008), pp. 87–96 (2008)
10. DMello, S.K., Craig, S.D., Witherspoon, A., McDaniel, B. and Graesser, A.: Automatic detection of learners affect from conversational cues. User Model User-Adap Inter, DOI 10.1007/s11257-007-9037-6 (2008)
11. D'Mello, S.K. and Graesser, A.: Language and Discourse Are Powerful Signals of Student Emotions during Tutoring. IEEE Transactions on Learning Technologies, Vol. 5(4), pp. 304–317, IEEE Computer Society (2012)
12. Moore, J.D., Tian, L. and Lai, C.: Word-Level Emotion Recognition Using High-Level Features. Computational Linguistics and Intelligent Text Processing (CICLing 2014), pp. 17–31 (2014)
13. Morency, L.P., Oviatt, S., Scherer, S., Weibel, N. and Worsley, M.: ICMI 2013 grand challenge workshop on multimodal learning analytics. In Proceedings of the 15th ACM on International conference on multimodal interaction (ICMI 2013), pp. 373–378 (2013)
14. Pardos, Z.A., Baker, R.S.J.D, San Pedro, M., Gowda, S.M. and Gowda, S.M.: Affective States and State Tests: Investigating How Affect and Engagement during the School Year Predict End-of-Year Learning Outcomes. Journal of Learning Analytics, Vol. 1(1), Inaugural issue, pp. 107–128 (2014)
15. Purandare, A. and Litman, D.: Humor: Prosody Analysis and Automatic Recognition for F * R * I * E * N * D * S *. In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP 2006), pp. 208–215 (2006)
16. San Pedro, M.O.C., Baker, R.S.J.D., Bowers, A. and Heffernan, N.: Predicting College Enrollment from Student Interaction with an Intelligent Tutoring System in Middle School. In Proceedings of the 6th International Conference on Educational Data Mining (EDM 2013), pp. 177–184 (2013)
17. Schuller, B., Batliner, A., Steidl, S. and Seppi, D.: Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. Speech Communication, Elsevier (2011)
18. Worsley, M. and Blikstein, P.: What's an Expert? Using Learning Analytics to Identify Emergent Markers of Expertise through Automated Speech, Sentiment and Sketch Analysis. In Proceedings of the 4th International Conference on Educational Data Mining (EDM '11), pp. 235–240 (2011)