Analysis of Modality-Based Presentation Skills Using Sequential Models

Su Shwe Yi Tun[1], Shogo Okada[1](✉), Hung-Hsuan Huang[2],
and Chee Wee Leong[3]

[1] Japan Advanced Institute of Science and Technology, Nomi, Japan
{sushweyitun,okada-s}@jaist.ac.jp
[2] The University of Fukuchiyama, Fukuchiyama, Japan
hhhuang@acm.org
[3] Educational Testing Service, Princeton, USA
cleong@ets.org

Abstract. This paper presents an analysis of informative presentations using sequential multimodal modeling for automatic assessment of presentation performance. For this purpose, we transform a single video into multiple time-series segments that are provided as inputs to sequential models, such as Long Short-Term Memory (LSTM). This sequence modeling approach enables us to capture the time-series change of multimodal behaviors during the presentation. We proposed variants of sequential models that improve the accuracy of performance prediction over non-sequential models. Moreover, we performed segment analysis on the sequential models to analyze how relevant information from various segments can lead to better performance in sequential prediction models.

Keywords: Social signal processing · Multimodal · Presentation skills · Sequence modelling

1 Introduction

Communication is one of the essential essences of human life. Verbal and nonverbal behaviors of human are used to predict the outcome of social interactions. Indeed, communication skills have been one of the important factors in affecting decisions in employment and other high-stakes situations. In the literature, there exists many studies that are focused on the training, feedback and assessment of communication skills, including those focused on monologue scenarios such as public speaking [4,20,25], business presentations [26] or social meeting [18], as well as those focused on communication skills in dyadic interaction situations, including the job interviews [5,16], group interactions [17,21] and human-computer interactions [10,23,24].

An oral presentation is a type of communication focused on a specific topic given to a potentiallylarge group of people. Intuitively, a good speaker should be articulate, organized, and purposeful to influence outcomes through the delivery

G. Meiselwitz (Ed.): HCII 2021, LNCS 12774, pp. 358–369, 2021.
https://doi.org/10.1007/978-3-030-77626-8_24

of the talk. While the success of presentation largely depends on the content of the talk, the speaker's verbal behavior, non-verbal (visual) cues, such as body language and gestures, also play a significant role. Nevertheless, good presenters can still adopt and practice presentation styles that differ from one another, resulting in significant challenges in modeling this variability in the assessment of presentations. To date, many studies focused on the automatic assessment of oral communication tasks [4,20,25] have relied on using features from various modalities to develop automatic assessment modelsto predict the scores assigned by human expert raters.

Automatic assessment of presentation skills can be performed using both verbal and non-verbal cues of the whole presentation or thin slices extracted from a video presentation [6,11], using different machine learning algorithms. With the exception of a few efforts [9,13], most efforts so far have relied on traditional machine learning approaches, as deep learning methods often require large amounts of *labeled* data for training, which is expensive and laborious to obtain for videos.

In our work, we use the time-series sequences from a dataset of 81 videos as input representation for training sequential models in an effort to do a comparative evaluation between sequential and non-sequential classifiers, such as SVM, when applied to the oral presentation assessment task. Sequential models, such as LSTM, are used to model a sequence of behavioral patterns over time during the presentation. Those information from time-series sequences can help improve presentation scoring accuracy. Evaluation on the presentation dataset shows that all scores of the proposed sequential models for each modality, except for the modality score of visual modality, significantly outperform non-sequential models with the best overall score (0.609) of audio modality using Stacked LSTM and best modality score (0.608) of text modality using RNN. Additionally, we analyze how each segment in a presentation can potentially contribute to improve the assessment accuracy of such presentations. Moreover, we discover which segment of presentation has more impact on the presentation assessment using sequential models and show that the presentation segments have different effectiveness to sequential models depending on the specific modality of a presentation.

2 Related Work

Previously, Haley et al. [14] collected an informative oral presentation dataset and described how information from each specific modality presented to a rater affects her judgment in the assessment of presentation tasks and investigated automatic assessment of presentation content using modality-specific machine learning features and model. They presented multi modal model prediction results on the dataset with overall and modality-score labels by fusing the modality-based features in an incremental approach (Text, Text + Audio, Text + Audio + Video). However, in their work, they did not investigate unimodal approaches as they mainly focus on the incremental fusion of modality-based features to investigate the presentation assessment improvement. They also did not explore regression modeling approaches from a deep learning perspective. In our work, we investigate each modality's effectiveness through sequential models to see how each modality affects the presentation.

In another line of work, Haider et al. [8] proposed an active data representation using audio-video segments of students' presentation and unsupervised self-organizing mapping for automatic scoring of delivery skills along with feedback generation. They described a data representation of videos using low-level audio descriptors and video descriptors (modeling body postures and movement). They created fusion models for those low-level descriptors to evaluate public speaking abilities. Additionally, they proposed a feedback method to flag presentation segments requiring improvement to users.

Kimani et al. [11] used HMM with state transition to provide feedback to the presenter and improve the presentation quality assessment results. They transform the overall presentation quality into states that represent the presenter's gaze, gesture, audience interaction, etc., and show how state-based representation improves the presentation results.

3 Data

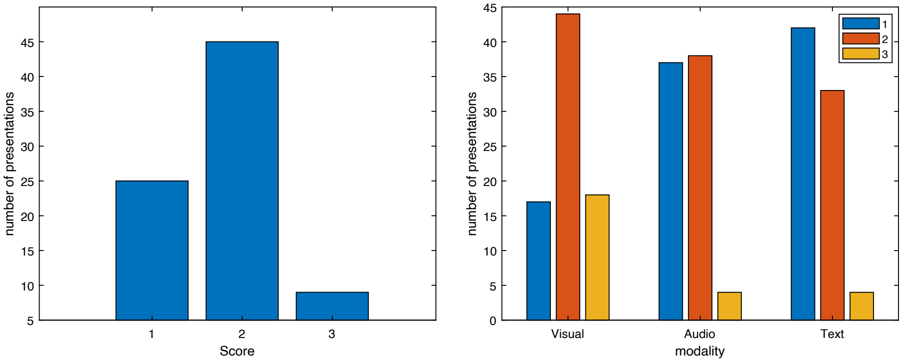
In this section, the English oral presentation dataset [14] is used in this study, and its annotations are described.

Task and Participants. The dataset contains videos of 81 college students from the United States giving an informative presentation for high school freshman students about what to prepare when choosing and applying to colleges. Note that the participants were asked to share their knowledge and information on college preparation instead of persuading the students to apply for college. The task involved (i) preparing a checklist to consider when selecting and applying to college, (ii) preparing for the presentation, and (iii) presenting an oral presentation for three minutes. The data were collected through participants interacting with HALEF [1] via a Web page, which is an open-source, cloud-based dialog system. In addition to the presentation, the participants also answered a few background information survey questions after the presentation recording.

Annotation. The annotation of the dataset was performed by human experts scoring on each presentation using an oral communication scoring rubric. Each presentation is scored on the content dimension of the rubric using a Likert scale of 0 to 4, where 0 is 'off-topic', '1' is deficient, '2' is 'weak', '3' is 'competent' and '4' is 'proficient'. Annotation is performed by two experienced raters for each of the three modalities, i.e., audio, video, and text. If there is a discrepancy in score level of more than one, then a third rater will be asked to perform the annotation. The raters provide three types of modality scores (audio, video, text) to each presentation. Two types of scores are defined for automatic assessment:

(1) overall score, which is the rounded median of all modality scores (audio, video, text) (2) modality-specific score (modality score) for each presentation, which is judged by observing only one modality (audio or video or text). The scores are assigned from the raters.

Scores Distribution. Due to the small dataset, we use the rounded down median of the two (or rounded median of three scores) as the final score for each presentation and combine the lowest two classes into a single class, which results in a three-class distribution. Figure 1 shows the distributions of both overall and modality scores of the oral presentation dataset.



(a) Score Distribution for overall score (b) Score Distribution for modality score

Fig. 1. Distribution of scores by (a) overall score (b) modality score

4 Multimodal Feature Extraction

Multimodal feature extraction of the dataset is performed automatically. We extracted acoustic information, facial expressions as a non-verbal aspect, and word-level features from the spoken utterances of the users as a verbal aspect. We extracted acoustic and visual features in an automatic manner and the word-level features are extracted using a cloud-based automatic speech recognition system. The following sections explain how these features are extracted.

4.1 Linguistic Features

Linguistic features are extracted from transcriptions. We extracted word embedding features for text computed using the word2vec [15] method. Firstly, we tokenized words from transcriptions and removed stop words using the Natural Language Toolkit library (NLTK) [3], and trained a word embedding model using the tokenized words via the Genism modeling toolkit [27]. The word2vec model projects our corpus with a vocabulary size of 1110 into the embedded vector space (embedding size of 200-D word2vec features). We converted each word in the transcription file into 200-D word2vec features and aggregated whole word embedding from each transcription into a single embedding input using sequential data modeling approach.

4.2 Acoustic Features

For acoustic modality, each audio file is first segmented into 5-s segments with an overlap of 1.5 s, and speech-based features are extracted using COVAREP [7]. The acoustic features set contains the prosodic features, voice quality information, and spectral information. Then, we computed the statistical values: mean, maximum, minimum, median, standard deviation, variance, kurtosis, skewness, percentile values for each feature and used them as acoustic features. Lastly, we combine all the segments of a given audio file into one feature vector, and feature selection is performed via the correlation matrix to select the top 100 features as the feature set for the model.

4.3 Visual Features

For video modality, each video file is also first segmented into 5-s segments with an overlap of 1.5 s and extracted time-series features at a sampling rate of 10 FPS using the OpenFace Toolkit [2]. We then used the 2D facial landmark data from eyes, mouth, eyebrows, and eye landmark data to calculate the velocity and acceleration of each data point and the mean value of the 18 facial AU features. Finally, we combine all the segments of a given file into one feature vector.

The transcription is annotated by timestamp per each utterance, which only contains timestamps of start and end time. Therefore, we cannot align the audio and video frame timing with each word in the transcriptions. Table 1 describe the details of features used in the experiments.

5 Experiments

In this section, the experiment is performed for each modality based on two labels: (i) overall score, and, (ii) modality score of that modality. In order to account for the variable length of input sequences, we used zero padding to normalize the length of the input sequence data. We evaluate both sequential and non-sequential models in two ways using (1) overall labels, and (2) modality-specific labels, where the labels specify performance levels based on the presentation content. For the experiments, we used the 81 samples that were obtained from the 81 participants.

Non-sequential Classification Models. For the non-sequential models: we experimented with two classification learners: Linear Support Vector Machine

Table 1. Summary of feature sets for presentation assessment

Modality	Feature names	Features
Linguistic	word2vec	200 dimension word2vec features
Audio	Prosodic	Fundamental Frequency(f0), voicing or not (VUV)
	Voice quality	Normalized Amplitude Quotient (NAQ), Quasi Open Quotient (QoQ), Amplitude difference between first two harmonics of the differential glottal source spectrum (H1H2), Parabolic Spectral Parameter (PSP), Maxima Dispersion Quotient (MDQ), Slope of Wavelet response (peakSlope), Shape parameter of LF glottal model (Rd), Detecting creaky voice (creak)
	Spectral	Mel-cepstral coefficient (MCEP 0–24), Harmonic model phase distortion mean (HMPDM 0–24), Phase distortion deviation (HMPDD 0–12)
Visual	2D facial landmarks	Four points from eyes, Four points from eyebrows, Four points around the mouth
	Facial action units	18 AU units

(LinearSVC) and Random Forest (RF). We use the average value of each element in a feature vector, which are extracted in Sect. 4, as an input to non-sequential models for each modality. We find the optimal hyperparameters of models using the grid search. We used SKLL[1], an open-source Python package that wraps around the sckit-learn package [19] for implementing the non-sequential learner.

Sequential Classification Models. For the sequential modeling approaches, we experimented with three models: RNN, LSTM, and Stacked-LSTM. An RNN model is composed of a single GRU layer with 128 units is used to extract

[1] https://github.com/EducationalTestingService/skll.

the features from input sequence data. The GRU layer was followed by a fully connected layer to learn RNN output. An output layer is used for predicting three labels. An LSTM model is composed of a single LSTM layer with 128 units is used to extract the features from the input sequence data, followed by a fully connected Layer to learn LSTM output. An output layer is used for predicting three labels. A Stacked-LSTM model is composed of two LSTM layers with 128 units is used to extract the features from the input sequence data, followed by a dropout (rate = 0.5) [22] layer. The LSTM layer was followed by 3 time-distributed dense layers to learn the LSTM output with 64, 32, and 16 for the number of units per layer, respectively. An output layer is used for predicting three labels. In our experiment, we used the Adam [12] optimizer with the learning rate of 0.001. For all models, sparse cross-entropy loss is used as the loss function with Softmax activation. We set the batch size to 16 and the number of epochs to 100. We used Keras with a TensorFlow backend for implementing the sequential models.

Evaluation Schemes. We experimented with each model using 10-fold cross-validation, and data normalization is performed using Z-normalization. We evaluated the experiments based on both average accuracy and balanced accuracy score since the dataset is imbalanced.

6 Experimental Results

Table 2 shows the comparison of average accuracy and balanced accuracy results of sequential models and non-sequential models. Since the dataset we used in this experiment was the imbalanced one, and we do not balance the data before training models in such results, we evaluated the results based on the accuracy score. We observed that except for the modality score of visual modality, all scores for other modalities achieved higher results in the sequential model than the non-sequential model. The highest accuracy for text, visual, and audio for the overall score using the sequential models were 0.590, 0.609, and 0.581, respectively, while the modality score yielded accuracies of 0.608, 0.593, and 0.497. Overall, the best accuracy (0.609) is achieved using the Stacked LSTM learner using audio modality, while the best accuracy for the text modality (0.608)is achieved using RNN.

7 Analysis of Specific Modality by Segments

In the previous section, we explored the sequential models using the full time-series data. However, the use of full data may contain irrelevant information for the presentation assessment. To address this issue, we performed manual segment analysis for finding which segments are relevant to the presentation assessment. Since we do not have the annotation for the individual segments, we define the label of all segment slices to be equal to the annotated score of the whole presentation. For this analysis, we extracted segments from the

Table 2. Experimental results for content presentation score using sequential and non-sequential models

		Overall score		Modality score	
Modality	Learner	Accuracy	Balanced accuracy	Accuracy	Balanced accuracy
Text	LinearSVC	0. 581	0. 346	0. 482	0. 327
	Random Forest	0. 528	0. 362	0. 435	0. 292
	RNN	**0. 590**	**0. 397**	**0. 608**	**0. 550**
	LSTM	0. 556	0. 376	0. 596	0. 527
	Stacked LSTM	0. 581	0. 375	0. 569	0. 511
Audio	LinearSVC	0. 491	0. 320	0. 492	0. 345
	Random Forest	0. 556	0. 342	0. 543	0. 379
	RNN	0. 441	0. 329	0. 487	0. 401
	LSTM	0. 491	0. 350	**0. 593**	**0. 562**
	Stacked LSTM	**0. 609**	**0. 514**	**0. 593**	0. 513
Visual	LinearSVC	0. 385	0. 290	0. 461	0. 354
	Random Forest	**0. 581**	0. 488	**0. 497**	**0. 414**
	RNN	**0. 581**	**0. 501**	0. 449	0. 375
	LSTM	0. 568	0. 443	0. 464	0. 390
	Stacked LSTM	0. 539	0. 408	0. 428	0. 363

presentation based on the transcription timestamps, which results in a total of three segments per presentation. We performed analysis on all three segments using the sequential models above. We train the models on (1) each segment, and, (2) all segments. Additionally, the evaluation is performed on the previous sequential models and the best score among the three models is used as a final score for analysis. The score is evaluated based on the balanced accuracy of the whole presentation. Using utterance-level timestamps, we used two segmentation approaches, described below, for comparative evaluation.

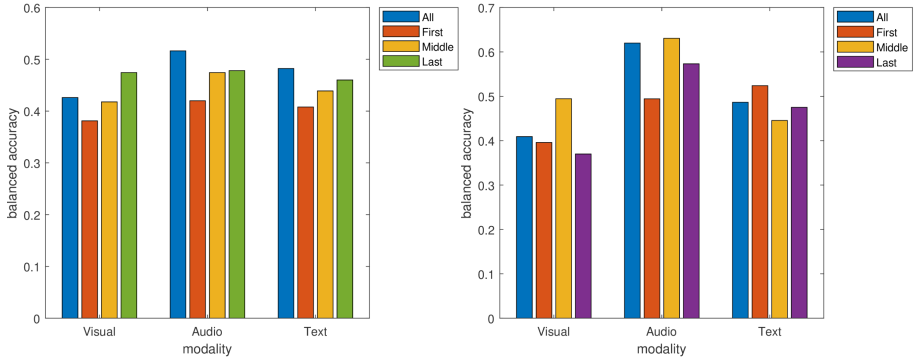
Nearest-Minute Segmentation. The data is segmented at the point closest to a minute-interval mark, which corresponds to the ending point of the last utterance in the previous segment, or the starting point of the first utterance in the next segment. The longest duration in the presentation dataset is three minutes two seconds while the shortest one is one minute and ten seconds. Because of this variable-length in presentations, we only have 79 last segments out of the initial 81 presentations.

Uniform Segmentation. The data is decomposed equally into first, middle, and last segments of equal lengths. The duration of each presentation may be different, but the duration of intra-presentation segments is always equal.

Results on Segment Analysis. Figures 2 and 3 describe the overall score and modality score for the nearest-minute segmentation and uniform segmentation approaches respectively. Table 3 shows the contributing segment that generated the best performance for each combination of modality and score metric. Regardless of segmentation approaches, we observe that the middle segments can be effective for predicting modality scores for both audio and visual modalities. Perhaps not so surprisingly, the first segment (representing first impressions) can be effective for predicting overall scores for both text and visual modalities when the uniform segmentation approach is used. Although we do not have exact annotation for each segment, the results show that not all segments are equal in their effectiveness as inputs to the models.

Table 3. Summary of segment contributions, by modality and score metric combinations

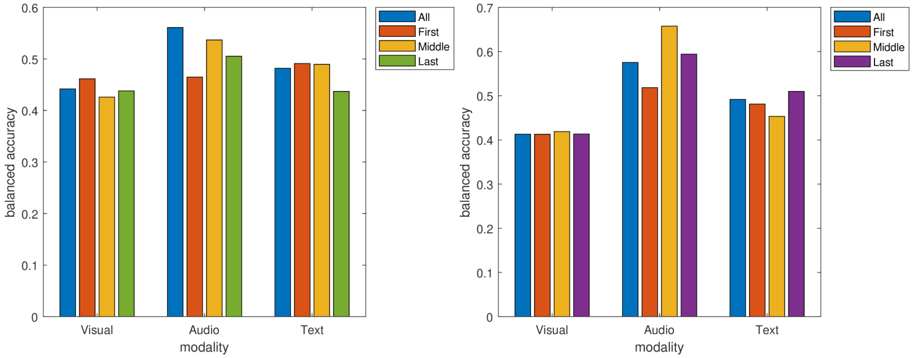
	Nearest-minute		Uniform	
Modality	Overall score	Modality score	Overall score	Modality score
Text	All	First	First	Last
Audio	All	Middle	All	Middle
Visual	Last	Middle	First	Middle



(a) Comparison of overall score between segments

(b) Comparison of modality score between segments

Fig. 2. Comparison of results obtained by nearest-minute segmentation



(a) Comparison of overall score between segments (b) Comparison of modality score between segments

Fig. 3. Comparison of results obtained by uniform segmentation

8 Discussions and Conclusion

In this work, we experimented with using unimodal approaches to model the performance of an oral presentation. While we did not explore any multimodal approach, we are mainly interested to employ sequential-based, unimodal models to study their efficacy against previously published approaches using non-sequential models on the same dataset. On this limited dataset of 81 videos, we achieved preliminary findings that sequential-based models are promising. As the next step, we plan to implement a fusion of the modalities using sequential models to predict human ratings on the same oral presentation task. In the future, we also plan to leverage transfer learning from a larger dataset to finetune models built for our small dataset. Moreover, given annotations on the more fine-grained, individual segments extracted from a video, it is possible that we further validate the effectiveness of our proposed segment-based, sequential approach on modeling oral presentations.

Acknowledgements. This work was partially supported by the Japan Society for the Promotion of Science (JSPS) KAKENHI Grant Numbers 19H01120, 19H01719 and JST AIP Trilateral AI Research, Grant Number JPMJCR20G6, Japan.

References

1. Halef. http://halef.org
2. Baltrusaitis, T., Zadeh, A., Lim, Y., Morency, L.: OpenFace 2.0: facial behavior analysis toolkit. In: Proceedings of the International Conference on Automatic Face and Gesture Recognition (FG), pp. 59–66 (2018)
3. Bird, S., Loper, E.: NLTK: the natural language toolkit. In: Proceedings of the ACL Interactive Poster and Demonstration Sessions, pp. 214–217. Barcelona, Spain (2004)

4. Chen, L., Feng, G., Joe, J., Leong, C.W., Kitchen, C., Lee, C.M.: Towards automated assessment of public speaking skills using multimodal cues. In: Proceedings of the International Conference on Multimodal Interaction (ICMI), pp. 200–203 (2014)
5. Chen, L., Zhao, R., Leong, C.W., Lehman, B., Feng, G., Hoque, M.E.: Automated video interview judgment on a large-sized corpus collected online. In: 2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 504–509. IEEE (2017)
6. Chollet, M., Scherer, S.: Assessing public speaking ability from thin slices of behavior. In: Procedings of the International Conference on Automatic Face and Gesture Recognition (FG), pp. 310–316 (2017)
7. Degottex, G., Kane, J., Drugman, T., Raitio, T., Scherer, S.: COVAREP: a collaborative voice analysis repository for speech technologies. In: Proceedings of the IEEE International Conference on Acoustics, Speech & Signal Processing (ICASSP) (2014)
8. Haider, F., Koutsombogera, M., Conlan, O., Vogel, C., Campbell, N., Luz, S.: An active data representation of videos for automatic scoring of oral presentation delivery skills and feedback generation. Frontiers Comput. Sci. **2**, 1 (2020)
9. Hemamou, L., Felhi, G., Vandenbussche, V., Martin, J.C., Clavel, C.: HireNet: a hierarchical attention model for the automatic analysis of asynchronous video job interviews. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 573–581 (2019)
10. Hoque, M.E., Courgeon, M., Martin, J.C., Mutlu, B., Picard, R.W.: MACH: my automated conversation coach. In: Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing, pp. 697–706. New York, USA (2013)
11. Kimani, E., Murali, P., Shamekhi, A., Parmar, D., Munikoti, S., Bickmore, T.: Multimodal assessment of oral presentations using HMMs. In: Proceedings of the International Conference on Multimodal Interaction (ICMI), pp. 650–654. New York, USA (2020)
12. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, 7–9 May 2015, Conference Track Proceedings (2015)
13. Leong, C.W., et al.: To trust, or not to trust? A study of human bias in automated video interview assessments. arXiv preprint arXiv:1911.13248 (2019)
14. Lepp, H., Leong, C.W., Roohr, K., Martin-Raugh, M., Ramanarayanan, V.: Effect of modality on human and machine scoring of presentation videos. In: Proceedings of the International Conference on Multimodal Interaction (ICMI), pp. 630–634 (2020)
15. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781 (2013)
16. Nguyen, L., Frauendorfer, D., Mast, M., Gatica-Perez, D.: Hire me: computational inference of hirability in employment interviews based on nonverbal behavior. IEEE Trans. Multimed. **16**, 1018–1031 (2014)
17. Okada, S., et al.: Estimating communication skills using dialogue acts and nonverbal features in multiple discussion datasets. In: Proceedings of the International Conference on Multimodal Interaction (ICMI), pp. 169–176. New York, USA (2016)
18. Park, S., Shim, H.S., Chatterjee, M., Sagae, K., Morency, L.P.: Computational analysis of persuasiveness in social multimedia: a novel dataset and multimodal prediction approach. In: Proceedings of the International Conference on Multimodal Interaction (ICMI), pp. 50–57. New York, USA (2014)

19. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
20. Ramanarayanan, V., Leong, C.W., Chen, L., Feng, G., Suendermann-Oeft, D.: Evaluating speech, face, emotion and body movement time-series features for automated multimodal presentation scoring. In: Proceedings of the International Conference on Multimodal Interaction (ICMI), pp. 23–30 (2015)
21. Sanchez-Cortes, D., Aran, O., Mast, M., Gatica-Perez, D.: A nonverbal behavior approach to identify emergent leaders in small groups. IEEE Trans. Multimed. **14**, 816–832 (2012)
22. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R.: Dropout: a simple way to prevent neural networks from overfitting. J. Mach. Learn. Res. **15**(56), 1929–1958 (2014)
23. Tanaka, H., et al.: Automated social skills trainer. In: Proceedings of the International Conference on Intelligent User Interfaces (IUI), pp. 17–27. New York, USA (2015)
24. Trinh, H., Asadi, R., Edge, D., Bickmore, T.: RoboCOP: a robotic coach for oral presentations. In: Proceedings of the ACM Interactive Mobile, Wearable and Ubiquitous Technologies 1(2) (2017)
25. Wörtwein, T., Chollet, M., Schauerte, B., Morency, L.P., Stiefelhagen, R., Scherer, S.: Multimodal public speaking performance assessment. In: Proceedings of the International Conference on Multimodal Interaction (ICMI), pp. 43–50 (2015)
26. Yagi, Y., Okada, S., Shiobara, S., Sugimura, S.: Predicting multimodal presentation skills based on instance weighting domain adaptation. J. Multimod. User Interfaces 1–16 (2021). https://doi.org/10.1007/s12193-021-00367-x
27. Řehůřek, R., Sojka, P.: Software framework for topic modelling with large corpora. In: Proceedings of LREC 2010 Workshop New Challenges for NLP Frameworks, pp. 46–50. Valletta, Malta (2010)