



StyleX: A Corpus of Educational Videos for Research on Speaking Styles and their Impact on Engagement and Learning

Harish Arsikere, Sonal Patil, Ranjeet Kumar, Kundan Shrivastava, Om Deshmukh

Data Analytics Lab, Xerox Research Center–India (XRCI), Bangalore, Karnataka, India

{Harish.Arsikere, Sonal.Patil, Ranjeet.Kumar2, Kundan.Shrivastava, Om.Deshmukh}@xerox.com

Abstract

In the context of learning through educational videos, the material chosen for a given topic must not only be relevant but also engaging to the consumer—ensuring better understanding and retention of content. This paper focuses on the speaking style of instructors, which is an important aspect driving student engagement. We present StyleX, a corpus of 450 1-minute video clips featuring 50 instructors, 10 topics in engineering and various accents of English. With the help of a large student population (304 in total), we study the impact of four speaking-style dimensions (liveliness, clarity, fluency and formality) on engagement and learning. Based on the in-classroom evaluations of 250 clips (> 20 simultaneous evaluators per clip), we find that liveliness and clarity are the most important dimensions (correlation with engagement and learning > 0.8), followed by fluency and formality. Familiarity with topics has a significant effect on the evaluators’ ratings, while the instructors’ accent and gender do not. StyleX represents the first large-scale effort of its kind in terms of the clip duration used and the number of topics, instructors and evaluators involved. This is also the first study, to our knowledge, on the explicit relationship between speaking style and engagement.

Index Terms: speaking style, educational videos, engagement, liveliness, clarity

1. Introduction

Online educational videos have become an important supplement to textual material (books, lecture notes, research articles, etc.), for self learning as well as classroom instruction [1, 2]. This is particularly true in the case of developing nations like India, where meeting the demand for quality classroom instruction can sometimes be a challenge. Given this significance of online learning, it is important to be able to mine the web for videos that are not only content appropriate (which is an obvious requirement) but also *as engaging as possible* to the consumer. As recent research has shown, student engagement leads to better understanding and retention of content while minimizing “in-video drop-out rates” [3, 4, 5, 6]. The present work focuses on the speaking style used by instructors, which is one of the main factors driving student engagement (the other factors being body language, audio quality, props used, etc.). The two main contributions of this paper are: (1) a sizable corpus of annotated educational videos (named StyleX) to enable research on speaking styles, and (2) preliminary insights into the impact of speaking-style dimensions on engagement and learning.

StyleX comprises of 450 one-minute video clips featuring 50 different instructors, 10 major topics in engineering and various accents of English. With the help of a large student population (304 in total), 250 clips have been subjectively evaluated

so far. StyleX represents the first large-scale effort of its kind in terms of the clip duration used for evaluation, the number of clips, topics, instructors and evaluators involved, and the controlled evaluation setup used to minimize inter-evaluator variability. The motivation for this data collection is two fold: (1) to uncover the dimensions of speaking style that contribute most to student engagement, and (2) to develop algorithms for the automatic ranking of instructional videos based on their ability to be engaging or based on a certain user-defined dimension (say fluency). The first goal is addressed here—laying the foundation to address the second goal in future studies.

Previous studies have provided various characterizations of speaking style. A few decades ago, Joos and Zwicky studied the dimension of casualness in conversational speech, and defined five degrees of casualness to describe speaking style [7, 8]. Eskenazi extended this to include two more dimensions: intelligibility and familiarity (with the listener) [9]. In the context of public speaking, Rosenberg and Hirschberg studied the elements of style that correlate with perceived charisma [10], while Strangert and Gustafson presented analyses to distinguish between good and bad speakers [11]. Along similar lines, the recent paralinguistic challenges at Interspeech have focused on the characterization and prediction of likability [12, 13].

In the context of educational instruction, only one previous study (to our knowledge) has attempted to characterize speaking style [14]. The authors of [14] used crowd-sourced descriptors of 100 video clips to identify six speaking-style dimensions, of which liveliness and speaking rate showed good inter-evaluator agreement. Using simple acoustic features and LASSO regression, the authors also developed automatic methods to predict liveliness and speaking rate. This paper is philosophically similar to [14], but novel in three important ways: (1) the size and diversity of the corpus collected (Section 2), (2) the methods employed to obtain evaluator feedback (Section 3), and (3) an explicit study of the relationship between speaking-style dimensions and engagement (Section 4). To encourage reproducible research in this important area, StyleX will be made publicly available towards the end of this year.

2. Video data

StyleX was compiled using videos downloaded from YouTube. We focused on undergraduate- and graduate-level topics in engineering given that a large amount of data (recorded video lectures) is readily available for this domain.

Three major engineering disciplines (Electrical, Mechanical and Computer Science) were chosen for our data collection, and 10 topics were covered in total (see row 1 of Table 1). Five unique instructors featured in the videos chosen for each topic, taking the sum total of instructors in the corpus to 50 (no instructor featured in more than one topic). A conscious effort

Topics	Data Structures and Algorithms, Optimization, Probability and Statistics, Machine Learning, Linear Algebra Signal Processing and Communication, Control Systems, Thermodynamics, Fluid Mechanics, Electronic Circuits
Schools	The Indian Institutes of Technology, Univ of New South Wales, MIT, Stanford Univ, Harvard Univ, Middle East Technical Univ, Duke Univ, Caltech, Univ of British Columbia, Univ of California at Berkeley, Princeton Univ, Univ of the West of England, Yale Univ, Univ of California at Irvine, Univ of Cambridge, Boston Univ, Colorado School of Mines, Indian Institute of Science
Accents	American (18), Asian (1), Australian (2), British (2), Greek (1), Indian (21), Middle Eastern (2), Spanish (1), Turkish (1), Unknown (1)

Table 1: Details of the topics, instructor affiliations and English accents (as determined by the first author) involved in StyleX.

was made to include instructors from schools across the globe (see row 2 of Table 1). Consequently, a wide range of English accents were also covered (see row 3 of Table 1). However, for the topics considered here, finding videos with female instructors was difficult (only 7 of the 50 instructors are females).

The video lectures downloaded from YouTube—3 per instructor, for a total of 150—were about 60 minutes in duration, on average. From each video lecture, we manually extracted 3 one-minute video clips (one each towards the beginning, middle and end of the lecture). The corpus therefore has 450 clips in total—9 per instructor, and 45 per topic. Whenever possible, we chose one “admin” clip per instructor in which there was little or no technical content. The rest of this paper is based only on the one-minute clips and not the hour-long video lectures. To maximize the amount of spoken content and also aid acoustic analyses in the future, we made sure to avoid long silences and “writing noise” (especially when it masked the speaker’s voice) while selecting the one-minute clips.

To allow evaluators a little more time before judging an instructor’s speaking style, the clip duration used here is higher than the duration used in [14] (20 seconds), [10] (2–28 seconds) and [11] (30–36 seconds). While some evaluators may need more than a minute to form a judgement (and not just an impression), we decided against a longer clip duration in the interest of securing a large number of evaluations without burdening the evaluators. Compared to the corpus in [14], StyleX also contains a larger number of clips (450 versus 318) corresponding to a larger number of topics (10 versus 2—high-school biology and chemistry).

3. Methods for subjective evaluation

The 450 one-minute video clips were binned into 18 “sessions” containing 25 clips each. The clip-to-session mapping was completely random except that no two clips in a given session were allowed to feature the same instructor.

3.1. Speaking-style dimensions and evaluation scheme

The authors of [14] used crowd sourcing to identify 6 education-specific dimensions of speaking style: liveliness, pleasantness, speaking rate, clarity, formality and confidence. They found that the first three dimensions were moderately correlated with one another, and that only a small proportion of the crowd-sourced descriptors alluded to the speakers’ confidence. Guided by these findings, the one-minute video clips in StyleX were subjectively evaluated on 4 dimensions of speaking style: liveliness, clarity, fluency (a dimension that has not been studied in the past) and formality. In addition, the clips were evaluated on their engagement levels and suitability for learning—the two ultimate goals of this study. While student engagement could be influenced by a variety of factors (e.g., the clarity of written content), the focus

of this work is solely restricted to speaking-style effects.

Given a video clip, the evaluators were asked to answer the following 8 questions (whose meanings were made clear prior to the task). They were asked to focus on the speaking styles without paying too much attention to the content.

1. Are you familiar with this topic? Response: Yes/No
2. Have you heard this speaker before? Response: Yes/No
3. How lively is the speaker?
Response: 0 (very dull) → 100 (very expressive)
4. How clearly is the speaker speaking?
Response: 0 (very unclear) → 100 (very clear)
5. How fluent is the speech?
Response: 0 (highly disfluent) → 100 (very fluent)
6. How formal is the speaker?
Response: 0 (very casual) → 100 (very bookish)
7. How engaging is this video?
Response: 0 (not at all) → 100 (highly engaging)
8. Would you prefer to learn from this lecturer?
Response: 0 (not at all) → 100 (absolutely)

The purpose of the first two questions was to check if evaluators would be influenced by their familiarity with certain topics or instructors, while the purpose of the last two questions was to study the effect of speaking-style dimensions on engagement and learning. Note that the last two questions capture different pieces of information. For instance, one may perceive a video to be engaging owing to the instructor’s liveliness, but unsuitable for learning owing to the use of informal language.

A 0–100 rating scale was chosen instead of the conventional 5- or 7-point Likert scale so that evaluators had more “freedom” while responding to the above questions. Additionally, a 0–100 scale was regarded as being better suited for correlation analyses. A small pilot study revealed that the potentially higher cognitive load imposed by the 0–100 scale was mitigated somewhat by the text descriptors at the extremes.

3.2. Modes of evaluation

Three avenues were explored for subjective evaluation: *classroom mode*, *full annotation mode* and *web mode*.

Classroom mode: This approach provides a controlled environment for the collection of real-time, simultaneous evaluations from sizable student groups. We visited five engineering colleges in Bangalore, and obtained student evaluations for a total of 10 StyleX sessions (250 video clips in total). The number of evaluators per session ranged from 21 to 43 (Table 2), and no evaluator participated in more than one session. All students were pursuing an undergraduate degree, and the majority of them were enrolled in Electrical, Computer Science or Mechanical Engineering.

For each evaluation session, students were asked to gather in their usual classrooms for about 60 minutes. Video clips were played to them with the help of a projector, a display screen and

speakers. The volume was adjusted until everyone in the classroom could clearly hear the instructor’s voice in a “test” video clip. The students were provided printed evaluation forms to write down their responses. They were given 30 seconds after playing each clip, and no clip was played more than once. The broadcasting setup (speakers, projector and display screen) varied from one session to another, but the environment within a session was identical across evaluators.

Full annotation mode: Our aim with this approach was to have 20 independent evaluators, each evaluating all 19 sessions of the StyleX database—offering an inherent calibration of the responses to some extent. Since this task is time intensive (completion time \approx 15 hours), we were able to secure only 6 complete evaluations at the time this paper was written.

In this mode, the entire database resides on the evaluator’s machine along with a tool that presents videos clips and records responses. Clips in sessions 1 through 19 are presented sequentially. An evaluation form (with the 8 questions described in Section 3.1) accompanies each clip; evaluators are also asked to justify their ratings on engagement and preference for learning (questions 7 and 8). The purpose of these additional responses is to uncover any hidden dimensions that may not already be modeled by questions 3–6. If the tool is closed midway through a session (i.e. before all 25 clips have been evaluated), it begins from clip 1 of that session upon relaunching. This ensures that at least 25 clips are evaluated in one sitting. Ten of the 20 evaluators have an engineering degree, while the remaining 10 evaluators are pursuing one.

Web mode: This approach provides access to a large number of evaluators with diverse backgrounds (not just engineering). The StyleX database was hosted on a server, and a web tool was designed to stream videos and evaluate them on any device with an internet connection. The tool is identical to its offline version in principle, except that it allows an evaluator to rate any number of videos as per his/her convenience. Each time an evaluator launches the tool, the video clip(s) to be presented are randomly chosen from the ones that haven’t been viewed before by him/her. The tool keeps track of the evaluators by assigning unique login names to them (when they launch the tool for the first time). The web page is currently live and user evaluations are being continuously gathered (these will be included in the StyleX distribution when it is made public).

4. Analysis of evaluator ratings

All the analyses below are based on the classroom evaluations that were obtained for 250 video clips.

4.1. Inter-evaluator agreement

The inter-evaluator agreement was computed on a session-by-session basis since the evaluator pool changed from one session to the next. To quantify inter-evaluator agreement, we used the average cross-correlation statistic (denoted by ρ_{ag}) [14]:

$$\rho_{ag} = \frac{1}{N} \sum_{j=1}^N \rho(x_j, \bar{x}_{-j}), \quad (1)$$

where $\rho()$, N , x_j and \bar{x}_{-j} denote Pearson’s correlation measure, the number of evaluators, ratings of the j^{th} evaluator, and averaged ratings of all except the j^{th} evaluator, respectively.

The ρ_{ag} values for questions 3–8 (Section 3.1) are shown in Table 2. Among the four speaking-style dimensions, liveliness and clarity show the highest inter-evaluator agreement on

average, followed by fluency and formality. The ρ_{ag} values for the first three dimensions are greater than 0.4 in most cases; this suggests good inter-evaluator agreement considering that each clip was rated by at least 20 evaluators. Also, the average inter-evaluator agreement for liveliness is consistent with the value reported in [14]. Formality shows particularly low values of ρ_{ag} , suggesting that the perception of this dimension is highly subjective. It is encouraging that the questions on engagement and learning also show good inter-evaluator agreement—comparable to that of liveliness and clarity.

4.2. Effect of topic familiarity, accent and gender

The evaluators in every session were familiar with some of the topics discussed in the video clips (based on their responses to question 1). Although they were asked to focus exclusively on the speaking styles, their ratings (on engagement and preference for learning) were higher, on an average, for topics that were familiar to them. The last two columns of Table 2 show that in 8 of the 10 sessions, the effect of topic familiarity is statistically significant ($p < 0.01$) as determined by a two-sample t-test. Similar trends were observed with regard to the evaluators’ preference for learning (results not shown). Given that a significant proportion of educational videos are used for learning unfamiliar topics, these results emphasize the importance of selecting videos that are as engaging as possible.

More than 40% of the instructors featuring in StyleX have an Indian accent. Since the evaluators were all Indian students, we checked if they had any accent preferences by performing two-sample t-tests (Indian-accented videos versus the rest) on their engagement ratings; no significant effects were found in any session. Similar t-tests were performed based on the instructors’ gender, and the results were found to favour the null hypothesis. These results might change as we get more diverse (and global) responses through the web evaluation mode.

4.3. A qualitative analysis of the ‘best’ and the ‘worst’

Table 3 lists the top 5 and bottom 5 instructors based on their average engagement ratings (column 7). The ratings of the top 5 instructors suggest that high engagement levels can result from being very lively, clear or fluent (as indicated by the values in bold face), while the ratings of the bottom 5 instructors suggest that low engagement levels are often caused by lack of liveliness (as indicated by the values in bold face). A post hoc analysis confirmed that the bottom 5 instructors indeed spoke with a flat tone. Column 4 of Table 3 suggests that an instructor’s ability (or inability) to be engaging through his/her speaking style is somewhat independent of the props used for instruction. Similarly, columns 5 and 6 suggest that engagement does not depend strongly on how videos are recorded (say, during a lecture with the instructor’s face shown). Indian-accented instructors appear at both the top and the bottom of the ranked list; this further substantiates the null hypothesis for accent effects (Section 4.2).

4.4. Speaking style versus engagement and learning

Table 4 presents the correlation coefficients between speaking-style dimensions on the one hand, and engagement and preference for learning on the other. The correlations were computed based on the average evaluator ratings of all 250 clips. All correlations are positive and statistically significant ($p < 0.01$), but liveliness and clarity emerge as the most important dimensions for engagement and learning, followed by fluency and formality. The correlations pertaining to formality, a dimension char-

Session #	Number of evaluators	Inter-evaluator agreement: ρ_{ag} , Eq. (1)						Average engagement rating	
		<i>liveliness</i>	<i>clarity</i>	<i>fluency</i>	<i>formality</i>	<i>engagement</i>	<i>learning</i>	<i>when topic is known</i>	<i>when topic is unknown</i>
1	36	0.55	0.47	0.50	0.30	0.43	0.46	53.9	49.8
2	43	0.46	0.44	0.01	0.21	0.41	0.41	55.2*	47.5
3	21	0.50	0.48	0.46	0.10	0.49	0.46	55.0*	46.5
4	26	0.39	0.44	0.46	0.22	0.48	0.47	63.1*	57.5
5	26	0.51	0.51	0.44	0.37	0.32	0.47	56.5*	43.3
7	30	0.47	0.47	0.44	0.09	0.45	0.46	60.6*	41.4
8	22	0.48	0.56	0.52	0.38	0.57	0.54	43.7	39.8
9	22	0.34	0.18	0.19	0.21	0.30	0.60	59.2*	49.0
10	40	0.52	0.50	0.49	0.35	0.45	0.49	55.0*	48.9
11	38	0.46	0.47	0.47	0.40	0.49	0.49	55.3*	41.0
Avg. across sessions:		0.47	0.45	0.40	0.26	0.44	0.49	55.7	46.5

Table 2: Classroom-mode results for inter-evaluator agreement and the effect of topic familiarity on engagement. Columns 3–8 show ρ_{ag} (Eq. (1)) values for questions 3–8 (Section 3.1). Asterisks in column 9 are used to indicate cases where the effect of topic familiarity is statistically significant ($p < 0.01$) as determined by a two-sample t-test. Note that session 6 has not been evaluated yet.

Rank	Gender	Accent	Instruction medium	In class?	Face shown?	Average evaluator ratings				
						<i>engagement</i>	<i>liveliness</i>	<i>clarity</i>	<i>fluency</i>	<i>formality</i>
1	M	Indian	white paper + pens	yes	yes	70.1	69.6	77.6	76.7	69.3
2	M	American	digital writing board	no	no	68.0	65.5	75.1	72.5	64.1
3	F	American	real-time animation	no	yes	65.7	65.3	72.9	74.8	67.9
4	M	Indian	blackboard + chalk	yes	yes	64.6	71.4	67.4	67.7	52.7
5	F	Indian	slide presentation	no	yes	61.7	62.0	71.9	70.5	68.1
46	M	British	slide presentation	yes	yes	41.2	44.9	46.9	50.6	54.1
47	M	Asian	blackboard + chalk	yes	yes	38.2	40.7	41.9	45.3	53.7
48	M	Indian	white paper + pens	no	yes	35.6	35.7	44.3	43.5	56.5
49	M	Arabic	slide presentation	no	no	32.9	34.1	43.1	44.2	51.2
50	M	Indian	slide presentation	no	yes	31.8	35.1	44.7	43.9	58.0

Table 3: Ranking of instructors based on their average engagement ratings. “In class?” is “yes” if the instructor’s videos were recorded during classroom lectures, and “Face shown?” is “yes” if the instructor’s face is visible. For instructors 1–5, values in bold face indicate dimensions with the highest ratings. For instructors 46–50, values in bold face indicate dimensions with the lowest ratings.

acterized by poor inter-evaluator agreement, must be interpreted with some caution. Not surprisingly, we found a strong correlation (0.89) between engagement and preference for learning.

Figure 1 shows box plots of the per-evaluator correlations between engagement and speaking-style dimensions (each correlation coefficient was computed using 25 pairs of ratings). The correlations pertaining to liveliness, clarity and fluency are strongly positive for most evaluators, while the correlations pertaining to formality show a large spread including values close to 0 as well as negative values. This suggests that formality is an unreliable dimension for making generic (not user-specific) recommendations of the most engaging videos on a given topic.

	Liveliness	Clarity	Fluency	Formality
Engagement	0.89	0.85	0.67	0.53
Learning	0.82	0.81	0.66	0.51

Table 4: Correlations between speaking-style dimensions, and engagement and preference for learning (computed based on all 250 clips evaluated through classroom mode).

5. Conclusion

StyleX is the largest corpus of its kind in terms of the number of instructional video clips (450), topics (10), instructors (50) and evaluators (at least 20 per clip) involved, and the clip duration (1 minute) used for evaluation. To address the diversity- and consistency-related issues that are typical of subjective evaluations, three methods of response collection were explored: *full* evaluation of the corpus, *in-classroom* evaluation of 25 clips at a time, and *web-based* evaluation of one or more clips.

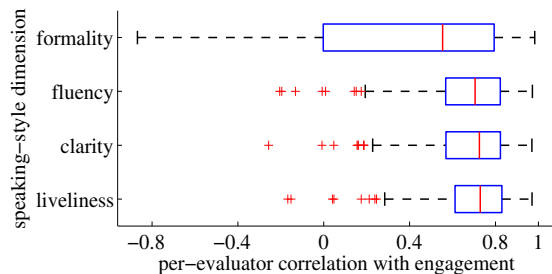


Figure 1: Box plots of the per-evaluator correlations between engagement and speaking-style dimensions.

In-classroom evaluation of 10 sessions shows that liveliness, clarity and fluency are important for engagement. Formality is found to be a highly-subjective dimension, and its impact on engagement is evaluator dependent. Topic familiarity has a significant effect on the perceived engagement levels, while the instructors’ accent, gender and the props used for instruction do not. The findings from this study will be leveraged to build intelligent systems that can model speaking styles and recommend the most engaging videos for a given topic. StyleX will be made public this year to further this important line of research.

6. Acknowledgements

We sincerely thank Prof. Marmar Mukhopadhyay for suggesting the classroom mode of evaluation, and Manjunath for compiling the student evaluations in digital format.

7. References

- [1] R. Agrawal, S. Gollapudi, K. Kenthapadi, N. Srivastava, and R. Velu, "Enriching textbooks through data mining," in *Proc. of the First ACM Symposium on Computing for Development*, 2010.
- [2] M. Miller, "Integrating online multimedia into college course and classroom: With application to the social sciences," *Journal of Online Learning and Teaching*, vol. 5, pp. 395–423, 2009.
- [3] P. J. Guo and K. Reinecke, "Demographic differences in how students navigate through MOOCs," in *Proceedings of the First ACM Conference on Learning@Scale*, 2014, pp. 21–30.
- [4] J. Kim, P. J. Guo, D. T. Seaton, P. Mitros, K. Z. Gajos, and R. C. Miller, "Understanding in-video dropouts and interaction peaks in online lecture videos," in *Proceedings of the First ACM Conference on Learning@Scale*, 2014, pp. 31–40.
- [5] E. Cutrell, S. Bala, C. Bansal, A. Cross, N. Datha, A. John, R. Kumar, M. Parthasarathy, S. Prakash, S. Rajamani *et al.*, "Massively empowered classroom: Enhancing technical education in India," Microsoft Research, Tech. Rep., 2013.
- [6] S. S. Krishnan and R. K. Sitaraman, "Video stream quality impacts viewer behavior: Inferring causality using quasi-experimental designs," *IEEE/ACM Transactions on Networking*, vol. 21, pp. 2001–2014, 2013.
- [7] M. Joos, "The isolation of styles," *Readings in the Sociology of Language*, pp. 185–191, 1968.
- [8] A. Zwicky, "On casual speech," in *Papers from the Eighth Regional Meeting of the Chicago Linguistic Society*, 1972, pp. 607–615.
- [9] M. Eskenazi, "Trends in speaking styles research," in *Proceedings of the Third European Conference on Speech Communication and Technology*, 1993, pp. 501–509.
- [10] A. Rosenberg and J. Hirschberg, "Acoustic/prosodic and lexical correlates of charismatic speech," in *Proceedings of Eurospeech*, 2005, pp. 513–516.
- [11] E. Strangert and J. Gustafson, "What makes a good speaker? subject ratings, acoustic measurements and perceptual evaluations," in *Proceedings of Interspeech*, 2008, pp. 1688–1691.
- [12] F. Burkhardt, B. Schuller, B. Weiss, and F. Wening, "Would you buy a car from me?-On the likability of telephone voices," in *Proceedings of Interspeech*, 2011, pp. 1557–1560.
- [13] B. Schuller, S. Steidl, A. Batliner, E. Nöth, A. Vinciarelli, F. Burkhardt, R. van Son, F. Wening, F. Eyben, T. Bocklet, G. Mohammadi, and B. Weiss, "The Interspeech 2012 Speaker Trait Challenge," in *Proc. of Interspeech*, 2012, pp. 254–257.
- [14] S. Mariooryad, A. Kannan, D. Hakkani-Tür, and E. Sriberg, "Automatic characterization of speaking styles in educational videos," in *Proceedings of ICASSP*, 2014, pp. 4848–4852.