Multimodal prediction of expertise and leadership in learning groups

Stefan Scherer
USC Institute for Creative Technologies
12015 Waterfront Dr.
Playa Vista, CA, USA
scherer@ict.usc.edu

Nadir Weibel
UCSD Dept. of Cognitive Science
9500 Gillman Dr.
La Jolla, CA, USA
weibel@ucsd.edu

Sharon Oviatt
Incaa Designs
11140 Wing Point Drive, N.E.
Bainbridge Island, WA, USA
sharon.oviatt@incaadesigns.org

Louis-Philippe Morency
USC Institute for Creative Technologies
12015 Waterfront Dr.
Playa Vista, CA, USA
morency@ict.usc.edu

ABSTRACT

In this study, we investigate low level predictors from audio and writing modalities for the separation and identification of socially dominant leaders and experts within a study group. We use a multimodal dataset of situated computer assisted group learning tasks: Groups of three high-school students solve a number of mathematical problems in two separate sessions. In order to automatically identify the socially dominant student and expert in the group we analyze a number of prosodic and voice quality features as well as writing-based features. In this preliminary study we identify a number of promising acoustic and writing predictors for the disambiguation of leaders, experts and other students. We believe that this exploratory study reveals key opportunities for future analysis of multimodal learning analytics based on a combination of audio and writing signals.

Keywords

Multimodal learning analytics, Writing, Speech

1. INTRODUCTION

The computational analysis of educational and learning situations such as investigations of individual student's performance, study group collaboration and multimedia assisted learning have gained a lot of momentum in recent years [1, 13, 6]. They investigate how students learn, struggle and improve their knowledge, with the goal in mind to optimize the acquisition of novel skills and to keep motivation up. One of the major interests is to foster expertise and to support weaker students. For example, we may want to challenge experts or not overwhelm weaker students. We aim at an on-the-fly analysis of nonverbal and written modality evaluation measuring expertise and motivation. An important challenge in this context is that socially dominant students might appear more knowledgeable than non-dominant students due to biases in social perception [11].

In this study, we aim to identify predictors of domain expertise that are distinct from behaviors associated with social dominance or leadership [9, 14]. As a secondary goal we analyze predictors of leadership in this context. We investigate low level predictors in both acoustic and written modalities. We analyze multimodal interactional data from a mathematical group learning task. Groups of three students interact and collaborate to solve several mathematical problems of varying difficulty levels. We analyze prosodic, voice quality, and turn taking features from the acoustic domain, such as articulation rate, pause durations, normalized amplitude quotient, and total fraction of speech and turns within the group interaction, to disambiguate between socially dominant, expert and other students. Further, we analyze geometric and timing related features within the written data, such as uninterrupted writing periods, strokes per second, fraction of overall time spent writing, and area covered in one uninterrupted writing period.

We chose these features as human nonverbal communication and prosodic measurements are found to be strong indicators of social dominance and other personality related phenomena [8, 14, 16]. Whereas, more planned and resource heavy behaviors such as the choice of spoken words, writing and sketching are expected to be more relevant for the identification of expertise [19].

We, therefore, hypothesize that nonverbal acoustic signal information will provide a more accurate predictor of social dominance in an individual, whereas writing and lexical/representational content will be more fertile predictors of expertise. For example, a socially dominant student will exhibit continuous speech floor-holding activity, but a dominant expert will write in a more structured way.

The remainder of the paper is organized as follows: We describe the used dataset in Section 2. We introduce and motivate the prosodic and written features in Section 3. These

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
MLA'12, October 26, 2012 Santa Monica, California, USA.
Copyright 2012 ACM 978-1-4503-1515-9/12/10 ...$15.00.

features then form the basis of our statistical analysis for which we report the results in Section 4. The results are discussed in Section 5 and Section 6 concludes the paper and introduces our future research directions.

2. DATASET

For this study we used an authentic and high-fidelity pilot corpus. It involves 12 sessions, with small groups of three students collaborating while solving mathematical problems. Data were collected on students' natural multimodal communication and activity patterns, including their speech, writing, facial expressions and physical movements. In total, approximately 15-18 hours of multimodal data is available during these situated problem solving and peer tutoring sessions.

2.1 Participants

Participants were 18 high-school students, that were organized in six same-gender groups of three varying in their performance capabilities (i.e., low- vs. high-performing). All students had just completed a course on the math content they were asked to solve. Students were motivated to understand how they arrived at each solution, because after completing each problem one student was randomly asked to explain their answer.

During each session, students engaged in authentic problem solving and peer tutoring as they worked on 16 geometry and algebra problems, four apiece representing easy, moderate, hard, and very hard difficulty levels. These levels were validated using: (1) teacher records of percent correct across a large student sample, and (2) students' percent correct solutions in previous studies. Each problem had a canonical correct answer. The following exemplifies a very hard problem: "Bart is paving a walkway in front of the library using a large steamroller. The steamroller's cylindrical wheel is 7 feet tall and 5 feet wide. How many square feet does a single revolution of the wheel cover?"

Each student group met for two sessions, during which students could view the math problems displayed one at a time on a tabletop computer screen. The sessions were facilitated by a tutoring system that could present the problems with accompanying visuals, terms and equations related to solving the problems, worked examples of how to solve problems, and the problem solutions. One student in the group was designated as the leader for a given session, and this designated leader switched on the group's second session to a different student. The student leader was responsible for interacting with the computer system that facilitated the group's problem solving.

2.2 Multimodal Data Collection

Natural multimodal data were recorded from 12 audio, visual, and pen signal streams. These included high-fidelity: (1) close-up camera views of each student from the waist up while working, a wide-angle contextual view, and a top-down contextual view of artifacts on the table; (2) close-talking microphone recordings of each students' speech, and one microphone of group discussion; (3) digital pen input for each student, using Anoto-based digital pens[1] and digital paper that streamed written input as students worked. All twelve media streams were time synchronized during data collection.

The multimodal nature and the numerous parallel streams of recorded information call for a tool to consolidate this data, framing it for meaningful analysis. To visualize the multiple streams of multimodal data and start to explore their relationship, we employed the ChronoViz software tool [18, 7]. ChronoViz integrates heterogeneous time-based data streams, enabling segmentation of the data and supporting analysis at different granularity levels, from data exploration to in-depth analysis of short moments of interest. While several systems have been developed to support various aspects of this analysis challenge, ChronoViz is unique in focusing on navigation of multiple diverse data sources, allowing researchers to visualize time-based data from multiple sources, interactively align data sources, navigate this data in flexible ways, and manually or automatically code the data with structured or unstructured text-based annotations. Annotations and derived data can also be created automatically through ChronoViz, by employing the provided analysis plugin framework that supports custom scripts.

As part of our analysis, we introduced a coding scheme for the mathematical problems, classifying them: (1) by difficulty level, (2) as correct or incorrect, (3) by which student initiated the answer. This establishes the ground-truth conditions for summarizing both problem-centric and student-centric expertise. To assess domain expertise associated with individual students, each students' cumulative problem-solving performance was calculated across their group's two sessions. When a student contributed an answer, the following numeric values were assigned per problem: total number of easy problems solved versus missed (+1 or -1 pt.), moderate problems solved versus missed (+2 or -2 pts.), hard problems (+3 or -3 pt.), and very hard problems (+4 or -4 pt.). Based on these totals, the expert student in each group was identified. In particular, the expert was identified as a the person in the group to have more than 10 points in both of the sessions. For one of the groups, no qualifying expert was found.

In the remainder of the paper assigned leaders will be referred to as L, experts as E and leading experts (i.e. when the assigned leader is also the expert of the group) as LE.

Figure 1 illustrates the richness of the collected data and the synchronized visualization of some of the raw dataset and the derived data that resulted from the analysis described in this paper.

3. ANALYZED FEATURES

In the following sections 3.1 and 3.2, we present the analyzed features from the audio and pen modalities. Features are motivated by related work. In particular, the speech features have been utilized in other studies investigating social signal related topics such as the identification of leadership and persuasiveness in political speeches [16].

3.1 Speech features

In this section we briefly discuss the prosodic features used in the statistical analysis. We chose these features as they have proven to be robust representatives of various prosodic phenomena in previous analysis. The features include the following measures:

- **Uninterrupted speech** is the duration of speech not

[1] http://www.anoto.com

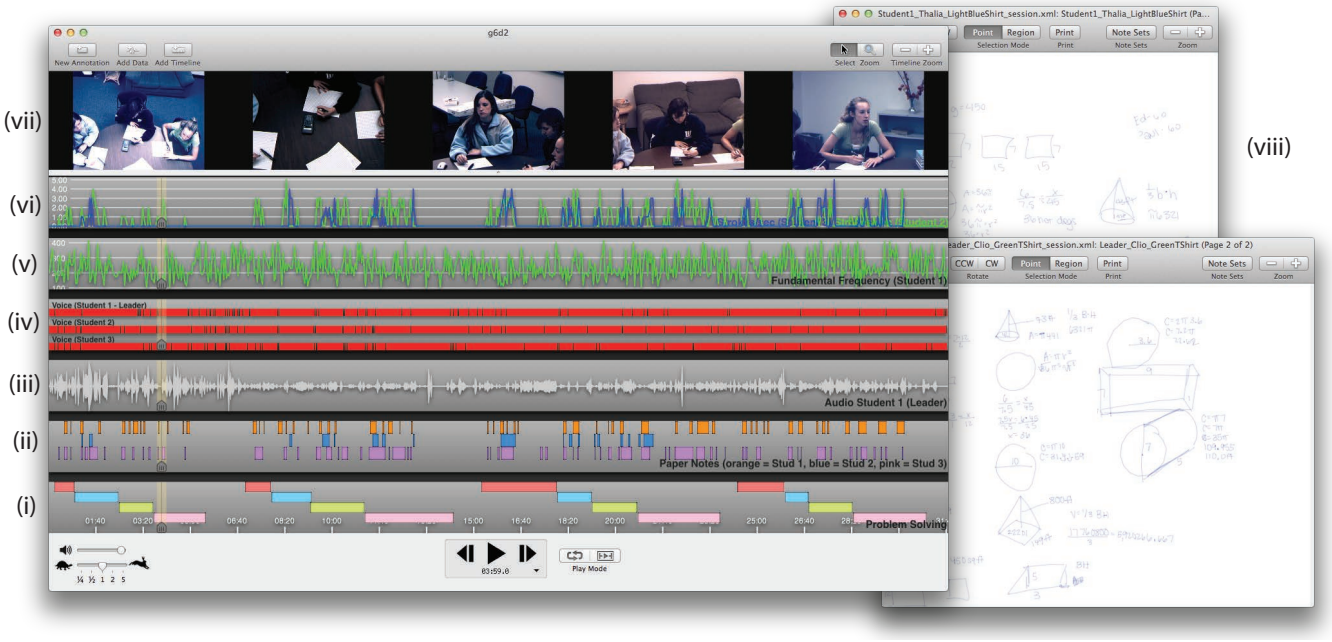

Figure 1: Multimodal dataset for one session visualized by ChronoViz. From bottom to top: (i) annotations identifying the different problems, (ii) time-based distribution of the digital paper notes, (iii) audio waveform of the speech signal, (iv) voice distribution, (v) fundamental frequency the audio signal, (vi) writing rate, (vii) video frames, (viii) digital representation of the paper notes.

interrupted by a pause greater than the $300ms$ threshold. Further, we investigate the total floor-time of the students as a fraction of the total duration of a session and the number of turns taken as the ratio of the total number of turns.

- **Pause duration** is a representation of the pauses over time, as well as the average time spent *not* speaking.
- **Energy (in dB)** is a measure of the intensity of the speech signal. Values closer to zero indicate louder speech.
- **Articulation rate** is calculated by identifying the number of syllables per second. The syllables are detected by identifying vowels in the speech.
- **Fundamental frequency (f_0)** is the base frequency of the speech signal. It is the frequency the vocal folds are vibrating at during voiced speech segments. f_0 is measured in semitones with a base frequency of 100 Hz.
- **Peak slope** is a measure suitable for the identification of breathy to tense voice qualities. Values closer to zero are considered as more breathy.
- **Spectral stationarity** is a value that captures the fluctuations and changes in the voice signal. High values indicate a stable vocal tract and little change in the speech (e.g. during a hesitation or sustained elongated vowels).

The following sections detail each acoustic feature.

Energy in dB

The energy of each speech frame is calculated on 32 ms windows with a shift of 10 ms (i.e. 100Hz sample rate). This speech window $w(t)$ is filtered with a hamming window and the energy

$$e(t) = \sum_{i=1}^{|w(t)|} w_i(t)^2 \tag{1}$$

is calculated and converted to the dB-scale

$$e_{dB}(t) = 10 \cdot \log_{10}(e(t)). \tag{2}$$

Articulation rate

Detection of syllable nuclei to calculate the articulation rate were made using the method introduced in [4], which is based on intensity peak detection of voiced segments of speech in Praat. On an abstract level, the script introduced in [4] follows a simple peak detection using the signal intensity. Peaks that are preceded and followed by considerable dips in intensity (based on a predefined threshold, e.g. -2 dB) are considered syllable nuclei candidates. All detected peaks, that are not voiced are removed from the selection in order to obtain the syllable nuclei without the need of a preceding transcription.

Fundamental frequency f_0

In [5], a method for f_0 tracking based on residual harmonics, which is especially suitable in noisy conditions, is introduced. The residual signal $r(t)$ is calculated from the speech signal $s(t)$ for each frame using inverse filtering. This process removes strong influences of noise and vocal tract reso-

nances. For each $r(t)$ the amplitude spectrum $E(f)$ is computed, showing peaks for the harmonics of f_0, the fundamental frequency. Then, the summation of residual harmonics (SRH) is computed as follows [5]:

$$SRH(f) = E(f) + \sum_{k=2}^{N_{harm}} [E(k \cdot f) - E((k - \frac{1}{2}) \cdot f)], \quad (3)$$

for $f \in [f_{0,\min}, f_{0,\max}]$, with $f_{0,min} = 50$ and $f_{0,max} = 300$. The frequency f for which $SRH(f)$ is maximal is considered the fundamental frequency of this frame. By using a simple threshold θ, the unvoiced frames are discarded as in [5].

Peak slope

This voice quality parameter is based on features derived following a wavelet based decomposition of the speech signal [12]. The parameter, named *peak slope*, is designed to identify glottal closure instances from glottal pulses with different closure characteristics. It was used to differentiate between breathy, modal, and tense voice qualities in [15]. The following equation is used for decomposing the speech signal:

$$g(t) = -\cos(2\pi f_n t) \cdot \exp(-\frac{t^2}{2\tau^2}), \quad (4)$$

where the speech signal $s(t)$ is convolved with $g(\frac{t}{s_i})$, and $s_i = 2^i$ and $i = 0, 1, 2, \ldots, 5$. This essentially is the application of an octave-band filter bank with the center frequencies being: 8 kHz, 4 kHz, 2 kHz, 1 kHz, 500 Hz and 250 Hz. Then the local maximum is measured at each of the signals obtained from the decomposition and a regression line is fit to these peaks. The peak slope parameter is the slope coefficient of the regression line, where higher negative values are associated with more tense voices and lower negative values with more breathy tones. In the original publication [12] this was carried out on individual phone segments. In the current study it is carried out on the frame level of the entire speech signal. A frame length of 32 ms and shift of 10 ms (i.e. 100Hz sampling rate) is used.

Spectral stationarity

To characterize the range of the prosodic inventory used over utterances, we make use of the so called *spectral stationarity* measure ss. This measurement was previously used in [17] as a way of modulating the transition cost used in the dynamic programming method used for f_0 tracking. Spectral stationarity, ss is measured with:

$$ss = \frac{0.2}{\text{itakura}(f_i, f_{i-k}) - 0.8} \quad \in [0, 1], \quad (5)$$

where itakura(.) is the Itakura distortion measure [10] of the current speech frame f_i and f_{i-k} is the previous frame with $k = 1$. We use a relatively long frame length of 60 ms (with as shift of 10 ms; sampling rate 100Hz) and frames are windowed with a Hamming window function before measuring ss. The long frame length was used in the attempt to characterize relatively long periods of maintained vocal tract articulation. ss is close to 1 when the spectral characteristics of adjacent frames are very similar and goes closer to 0 if the frames show a high degree of difference.

3.2 Writing features

The use of digital pens during data collection allows us to investigate basic writing features in naturalistic use of pen and paper. Writing features are defined by taking into account the timestamped strokes recorded by the digital pen as well as the distribution in space of the strokes.

- **Writing Rate** is calculated by identifying the number of strokes per second. A stroke is defined by a continuous sequence of writing activity between a *pen-down* and a *pen-up* event. To calculate writing rate 1 sec. windows are used and strokes are fitted in the window depending on the timestamp of the central point of the stroke.
- **Writing Area** is the area covered by a chunk of written text in mm^2. This is composed by the width and height of the bounding box enclosing the single chunk (or stroke)
- **Aspect Ratio** is the relationship between the height and the width of a chunk of written text (or a single stroke).
- **Pressure** is the force applied during the writing of the single strokes calculated in a scale from 1 to 255. This value is returned by a pressure sensor integrated in the digital pen, however from our experience, the distribution of the pressure values is not proportional to the force applied, resulting in a binary (or at most three-level) description of the pressure (hard, medium, soft).

Through further analysis of the temporal dynamics of the strokes and the chunk of written information, it is possible to extract higher level features describing the writing activity over time. This relies on the definition of a pause threshold that determine when subjects are moving from an entity of written information to another. Cheng and Rojas-Anaja [2] studied pause thresholds for written text in the context of geometrical shapes, numbers, and familiar words in terms of inter-chunk ($620ms$, $440ms$, $400ms$) and intra-chunk ($410ms$, $280ms$, $270ms$) distribution. Given the heterogeneous nature of our dataset we use a combined threshold for inter-chunk pause of $500ms$.

- **Uninterrupted writing** is the duration of sequential strokes not interrupted by a pause greater than the $500ms$ threshold. Additionally, we calculate the total writing time of the students as a fraction of the total duration of a session.
- **Pause distribution/Average pauses** is a representation of the pauses over time, as well as the average time spent *not* writing.

Further processing of the raw data enables to calculate derived features based on character, shape, or gesture recognition. This analysis requires the usage of dedicated algorithms and external software (e.g. intelligent character recognition[2], gesture recognition[3])to further process raw data. By applying those algorithms, we plan in the future to be able to recognize following features and integrate them in our analysis:

[2]E.g. MyScript ICR: http://www.myscript.com
[3]E.g. iGesture: http://www.igesture.org, $1 recognizer: http://depts.washington.edu/aimgroup/proj/dollar

	Speech	Pause	Art.	f_0 var.	Peak Slope	Energy	Stat.
L	.59 (.72)	1.36* (3.88)	4.11** (1.71)	4.16 (1.74)	-.07** (.16**)	-66.72** (17.32)	.09** (.13*)
NL	.58 (.79)	1.67* (4.96)	3.57** (1.60)	4.26 (1.69)	-.01** (.09**)	-74.20** (16.06)	.13** (.15*)
E	.58 (.79)	1.22 (3.76)	3.96 (1.64)	4.16 (1.64)	-.07* (.13)	-67.44 (16.91)	.08 (.13)
NE	.59 (.78)	1.56 (4.69)	3.88 (1.64)	4.26 (1.78)	-.01* (.10)	-71.56 (16.39)	.11 (.15)
LE	.61 (.98)	1.18 (3.69)	3.96 (1.70)	4.50 (1.58*)	-.15* (.20)	-65.06* (17.47)	.08 (.13)
L	.57 (.72)	1.36 (4.00)	4.28 (1.71)	4.14 (1.83*)	-.05* (.15)	-67.30 (17.13)	.10 (.13)
E	.58 (.78)	1.44 (5.16)	3.98 (1.59)	4.08 (1.68)	-.06* (.12)	-71.68* (16.86)	.10 (.14)

Table 1: Statistics and comparison of prosodic values for the different speakers denoted by L for leader, NL for non-leader, E for expert, NE for non-expert, and LE for leading expert in the respective analysis set. Both the median values for the respective feature and the median values of the standard deviations (in brackets) are displayed. Several significant differences between L and NL are found. Additionally, E and NE can be distinguished using the voice quality parameter peak slope. For the three group disambiguation we find peak slope and energy features to be useful.

- **Writing type** is the type of information that has been written. It can be textual or graphical information.
- **Text Style** is the style used to write textual information: cursive, block capitals, etc.
- **Characters** are the characters recognized by the used handwriting recognition engine.
- **Shapes and Gestures** are the geometrical figure recognized by a gesture recognizer. Shapes are static representations, while gestures encode also the temporal dynamics made available by the timestamped information recorded by the digital pen.

4. STATISTICAL EVALUATION

For the statistical evaluation we compared the median and standard deviation values for both speech and writing features between multiple groups of students. The groups are separated into leaders (L), non-leaders (NL), experts (E), non-experts (NE) and leading experts (LE) as described in Section 2. In total we conducted three separate sets of evaluation: (i) L vs. NL; (ii) E vs. NE and (iii) LE vs. E vs. L. We conducted independent t-tests and the significant results are reported in the respective tables.

4.1 Prosodic evaluation

The results for the prosodic evaluation are found in tables 1 and 2 for all eight analyzed features and the total speaking fraction and overall turn ratio. In the following we will highlight a couple of significant results for the respective evaluations.

(i) Leaders vs. Non-Leaders: There are multiple prosodic features that are significantly different for the assigned leaders and other students: The median pause duration for L is significantly lower (1.36 sec.) than that of NL (1.67 sec.) with $p < .05$. The distribution of the pause durations is visualized in Figure 2 (ii). Further, the articulation rate (Art. in Table 1) is significantly higher for L (4.11 syllables per sec.) than for NL (3.57 syllables per sec.) with $p < .01$. The distributions of the observations for articulation rate is visualized in Figure 2 (i). The median and variations of the peak slope parameter are also highly significantly different (-.07 (.16) for L vs. -.01 (.09) for NL) with $p < .01$. Furthermore, the measured spectral stationarity (Stat. in Table 1) is significantly lower for L (.09) than for NL (.13) with $p < .01$. The distribution of the observations for spectral stationarity are shown in Figure 2 (iii). The total speaking fraction of L is strongly significantly higher (.39) than that of NL (.31) with $p < .01$ - note that ratios within a group add up to more than one as there might be overlapping speech. Additionally, the ratio of turns (i.e. the fraction of a single person of all turns) is significantly higher for L (.22) than for NL (.19) with $p < .01$.

(ii) Experts vs. Non-Experts: For the analysis between E and NE the only observed significant difference is for the peak slope parameter with -.07 for E and -.01 for NE. The statistically not significant observations or articulation rate, pause duration and spectral stationarity are found in Figure 2 (i)-(iii).

(iii) Leader-Experts vs. Experts vs. Leaders: For the analysis between the groups LE, E and L we compare the values of LE the others in two separate independent t-tests. We found a significant difference between the f_0 variations of LE (1.58) and L (1.83) with $p < .05$. Additionally, the peak slope values for LE (-.15) are significantly different to both E (-.06) and L (-.05) with $p < .05$. The distribution of the values is shown in Figure 4. Further, the observed loudness is significantly higher for LE (-65.06 dB) than for E (-71.68 dB) with $p < .05$. The distribution of the values is shown in Figure 3. The total speaking fraction is significantly higher for LE (.45) than for E (.34).

4.2 Writing evaluation

We analyzed the main basic writing features described

	Total Speaking	Turn ratio	Total Writing
L	.39** (.08)	.22** (.03)	.09 (.03)
NL	.31** (.08)	.19** (.05)	.09 (.04)
E	.37 (.07)	.24 (.05)	.05 (.04)
NE	.31 (.09)	.20 (.05)	.10 (.04)
LE	.45* (.06)	.26 (.03)	.07 (.03)
L	.36 (.08)	.22 (.03)	.10 (.03)
E	.34* (.05)	.20 (.05)	.04 (.05)

Table 2: Statistics and comparison of total ratios of speaking fraction, total turn fraction, and total ratio of time spent writing values for the different student groups. Total speaking fraction and turn ratio are statistically significant predictors for L and NL. Total speaking fraction is significantly different for LE and E.

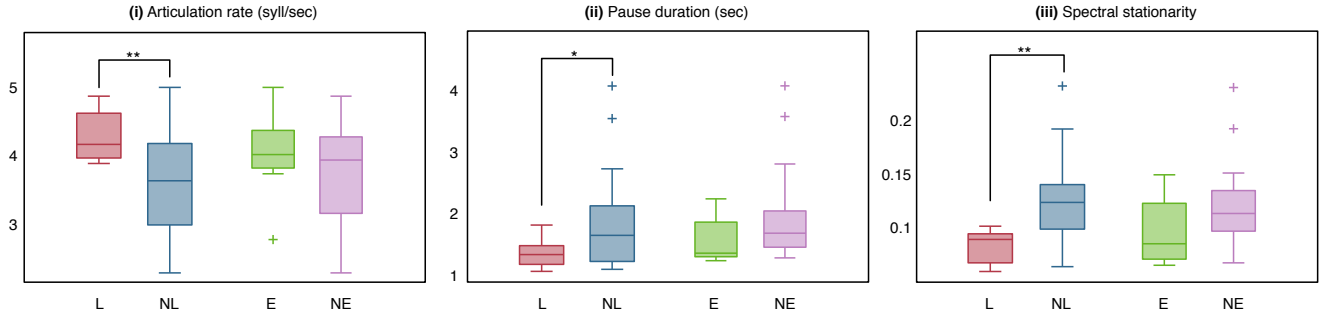

Figure 2: Statistical evaluation of three prosodic features (i)-(iii). Data is grouped in three: leaders (L), non-leaders (NL), experts (E) and non-experts (NE). Significant differences are marked with annotated brackets - * for $p < .05$ and ** for $p < .01$ in independent t-tests. Significant differences for all three parameters are found between L and NL. No significant variation is found between E and NE.

earlier, as well as the total writing fraction. Results of our analysis are highlighted in Table 3 and Table 2 respectively. In the remainder of this section we highlight the most important results emerging from this preliminary analysis. In the next section we then discuss the impact of these results on multimodal learning analytics.

(i) Leaders vs. Non-Leaders: As illustrated in Table 3, uninterrupted writing intervals are significantly shorter ($p < .05$) for L than for NL (L 1.20 sec. vs. NL 1.38 sec.). This trend is interestingly coupled with the pause intervals that highlight how L make average longer pauses (2.86 sec.) than NL (2.25 sec.). Even though this result does not show high significance, we believe that by expanding the dataset and carefully filtering data depending on the type of written text (see previous section's description of derived data) in the future this trend might be of significant outcome. In terms of the area covered, our analysis does not show significant difference between L (64.73 mm^2) and NL (81.36 mm^2). This is mostly due to a large standard deviation that might be adjusted in the future by taking into account derived features.

(ii) Experts vs. Non-Experts: Here, the average pause duration shows important differences between E (1.95 sec.) and NE (2.36 sec.). E and NE can also be distinguished by looking at the average ratio spent writing by E, which is only half as that of NE (E .05 vs. NE .10). Even though this result only approaches statistical significance, we feel that this could be a good predictor of expertise.

(iii) Leader-Experts vs. Experts vs. Leaders: With respect to comparing the three groups we found a couple of interesting results. As seen in Table 3, the average pause duration of LE (1.70 sec.) is smaller than that of L (2.95 sec.) and E (2.20 sec.). Further, the area covered on average is much larger for E (94.19 mm^2) than for LE (61.52 mm^2) and L (69.95 mm^2). Finally, also in this case the ratio of

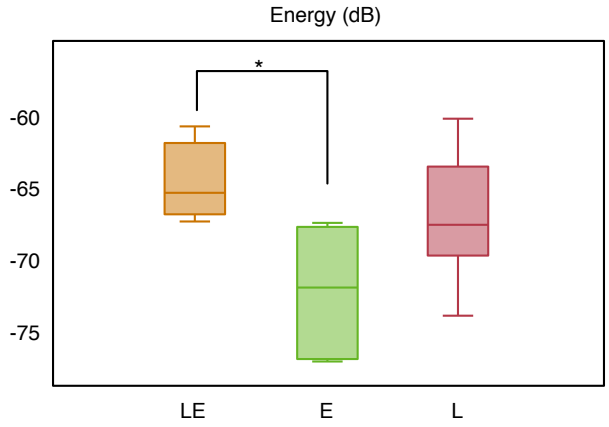

Figure 3: Statistical evaluation of energy (in dB) parameter. Data is grouped in three: leading experts (LE), experts (E) and leaders (L). Energy is significantly lower for E than for LE.

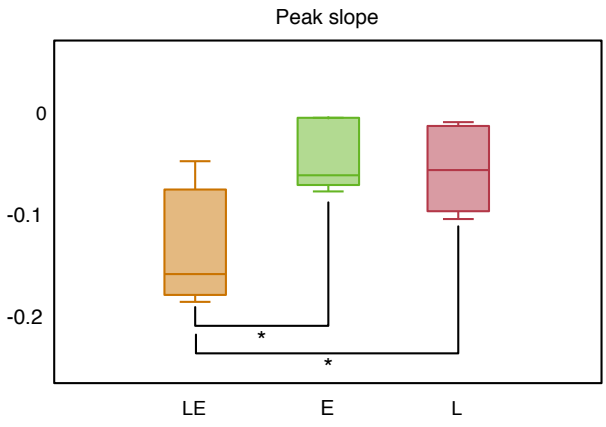

Figure 4: Statistical evaluation of peak slope parameter. Data is grouped in three: leading experts (LE), experts (E) and leaders (L). Peak slope is the only parameter in the study that shows statistical significant differences between LE and both other groups E and L. Tense voice quality is a strong predictor for the group LE.

	Writing	Pause	Area	Strokes
L	1.20* (1.25)	2.86 (101.26)	64.73 (346.97)	1.65 (.77)
NL	1.38* (1.43)	2.25 (120.96)	81.36 (395.29)	1.69 (.84)
E	1.28 (1.28)	1.95 (128.32)	78.61 (314.27)	1.69 (.81)
NE	1.32 (1.41)	2.36 (112.32)	77.69 (401.68)	1.69 (.84)
LE	1.19 (1.17)	1.70 (204.33)	61.52 (315.25)	1.65 (.76)
L	1.26 (1.28)	2.95 (100.36)	69.95 (385.61)	1.67 (.79)
E	1.39 (1.35)	2.20 (115.97)	94.19 (288.33)	1.70 (.83)

Table 3: Statistics and comparison of writing feature values for the different student groups. Uninterrupted continuous writing is significantly different between L and NL. Other trends are emerging, which are not significant.

total time spent writing is much lower for E (.04) than for L (.10) and LE (.07) (see Table 2).

While a lot needs to be explored in this triadic analysis, the emerging trends could be useful to frame future investigations.

5. DISCUSSION

In this section we discuss the findings and implications of the previous section in detail. Given the exploratory nature of our study, note that these are preliminary results based on a relatively small sample size. We would like therefore to regard the results as trends rather than as final.

As hypothesized, we found a number of significant predictors in the prosodic domain for the separation of the assigned leaders from other students and less predictors for the separation of experts and other students. In particular the articulation rate is higher and pause durations are significantly lower for leaders than for other students. This finding indicates that leaders tend to speak faster and take the floor more often than other students. This is further confirmed by the significantly higher speaking fraction and overall turn ratio reported in Table 2. It becomes apparent that the socially dominant leaders clearly take the floor more often than the other students and especially leader-experts tend to dominate the learning sessions clearly. This dominance might also be a byproduct of their roleplaying behavior in the situated learning task, as they were not necessarily socially dominant per se but rather the assigned leader of the group for the learning session.

Additionally, the peak slope parameter is significantly different between leaders and other students. This finding indicates that overall the voice quality of leaders is more tense than that of other students, which is in accordance with findings of other social signal research that indicates that socially dominant and persuasive speakers do use more tense voices than others [16]. Similarly, the observed speech energy is significantly higher for leaders than for other students, which is another indication of social dominance. Lastly, the monotony of the speech is significantly lower for leaders than for other students.

For the separation between experts and other students the only observed predictor that is significant is the peak slope parameter. This parameter additionally, shows significant differences in the three group analysis of leading experts, leaders and experts (see Figure 4). It is observed, that leading experts (i.e. the assigned leader is also the group's expert) are speaking with a significantly more tense voice than the others. A further significant predictor between leading experts and experts is the speech energy (see Figure 3). It is significantly higher for leading experts, which strengthens the assumption that energy is a strong factor of displaying leadership in the situated learning task.

Although many different trends emerged from the analysis of the basic writing features, only one of them is a significant predictor for leadership. Namely, the average duration of uninterrupted writing for leaders is shorter than for non-leaders, which could signify that leaders tend to spend more time in creating more elaborate representations of the problems stated. However, we believe that the other emerging trends are interesting and important for framing further analysis. In particular, the average area covered by an uninterrupted writing sequence is larger for non-leaders than for leaders. Further, experts use a much larger area within an uninterrupted writing sequence when compared to leader-experts and leaders. This could possibly signify that they are sketching larger schematic geometric representations of the present mathematical problem while not being busy instructing (i.e. leading) others. Additionally, the ratio of the total time spent writing by experts is only half as much as for non-experts. An even stronger effect for the ratio of the total time spent writing is found when comparing experts to leaders and leader-experts. This is an indication that experts require less time for the written parts of problem solving.

With this study we are setting the scene for the integration of written features in the domain of multimodal learning analytics. Although we couldn't find strong significant results, we believe that the reported trends are a good starting point for further investigation. We are currently developing more sophisticated and purpose tailored features. We further believe, that by combining the potentially powerful predictors resulting from the writing analysis with the more solid outcomes of the speech analysis, we might be able to effectively predict learning behavior in the future.

6. CONCLUSIONS AND FUTURE WORK

In this study we analyzed several low level predictors from the acoustic and written modalities to disambiguate leaders and experts in a computer assisted group learning dataset. We found a number of strong predictors for the identification of leaders or socially dominant students. For the identification, of experts from the speech domain we could identify the voice quality parameter peak slope to be a significant predictor. When comparing leader-experts (i.e. the expert and leader of a group coincide) with the rest of the leaders and experts, the peak slope parameter was found to be significantly different. This finding is very promising and we plan to further investigate this in the future to disambiguate the groups.

Within the written domain, we could identify a couple of interesting trends to disambiguate between experts and the rest of the students. Our findings indicate, that experts strongly plan what they write beforehand and write faster than other students. Further, the type of strokes or writing they produce (i.e. symbolic writing and sketching vs. prosaic writing) might be very different and should be investigated in future studies with more sophisticated writing based features. Overall, we could partially confirm our previous hypothesis that social dominance is assessable

using nonverbal and turn taking behaviors. Some promising trends based on written features show that expertise is predictable in combination with voice quality related features.

As future work we are planning to incorporate time dependent and turn-management related features similar to the steady conversational period evaluation in [3]. While in this preliminary study, we investigate the two modalities in a parallel fashion we expect an intertwined analysis of the two modalities to reveal stronger predictors for the disambiguation of expertise and leadership in the future. Further, we will start incorporating the visual data and will look at typical gaze behavior and movement patterns. We plan to investigate human performance and perception strategies to disambiguate experts and socially dominant students in user studies to confirm some of the measures and identify novel parameters. We anticipate that the use of an exploratory visualization and analysis tool such as ChronoViz will facilitate our initial analysis. The further integration of enhanced metrics for audio and written data based on the ones presented in the paper, could be the key to answer key questions for multimodal learning analytics.

Overall, the investigated data show promising results that motivate us to pursue the topic of leadership and expertise prediction in the future. The automatic prediction then can be used to improve the objective evaluation of group learning tasks and optimize individualized student specific learning strategies.

7. REFERENCES

[1] R. Baker and K. Yacef. The state of educational data mining in 2009: A review and future visions. *Journal of Educational Data Mining*, 1:3–17, 2009.

[2] P. Cheng and H. Rojas-Anaya. Measuring mathematic formula writing competence: An application of graphical protocol analysis. In *Proceedings of the Thirtieth Annual Conference of the Cognitive Science Society*, pages 869–874, 2008.

[3] M. Cristani, A. Pesarin, C. Drioli, A. Tavano, A. Perina, and V. Murino. Generative modeling and classification of dialogs by a low-level turn-taking feature. *Pattern Recognition*, 44(8):1785–1800, 2011.

[4] N. H. De Jong and T. Wempe. Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior research methods*, 41(2):385–390, 2009.

[5] T. Drugman and A. Abeer. Joint robust voicing detection and pitch estimation based on residual harmonics. In *Proceedings of Interspeech 2011*, pages 1973–1976. ISCA, 2011.

[6] S. Finkelstein, S. Scherer, A. Ogan, L.-P. Morency, and J. Cassell. Investigating the influence of virtual peers as dialect models on students' prosodic inventory. In *Workshop on Child, Computer and Interaction (WOCCI'12)*. ISCA, 2012.

[7] A. Fouse, N. Weibel, E. Hutchins, and J. D. Hollan. ChronoViz: A System for Supporting Navigation of Time-coded Data. In *Proceedings of CHI 2011 Interactivity Track, ACM Conference on Human Factors in Computing Systems*, Vancouver, Canada, May 2011. `http://chronoviz.com`.

[8] D. Gilbert. Thinking lightly about others: Automatic components of the social inference process. In J. Uleman and J. Bargh, editors, *Unintended Thought*, pages 189–211. Guilford Press, 1989.

[9] H. Hung, Y. Huang, G. Friedland, and D. Gatica-Perez. Estimating dominance in multi-party meetings using speaker diarization. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):847 –860, 2011.

[10] F. Itakura. Minimum prediction residual principle applied to speech recognition. *IEEE Transactions on Acoustics, Speech and Signal Processing*, ASSP-23:67–72, 1975.

[11] D. Kahneman. *Thinking, Fast and Slow*. Farrar, Strauss and Giroux, 2011.

[12] J. Kane and C. Gobl. Identifying regions of non-modal phonation using features of the wavelet transform. In *Proceedings of Interspeech 2011*, pages 177–180. ISCA, 2011.

[13] S. L. Oviatt. *Multimodal interfaces, The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*. Taylor and Francis, 3rd edition, 2012.

[14] A. Pentland. *Honest Signals - How they shape our world*. MIT Press Cambridge, 2008.

[15] S. Scherer, J. Kane, C. Gobl, and F. Schwenker. Investigating fuzzy-input fuzzy-output support vector machines for robust voice quality classification. *Computer Speech and Language*, 2012.

[16] S. Scherer, G. Layher, J. Kane, H. Neumann, and N. Campbell. An audiovisual political speech analysis incorporating eye-tracking and perception data. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, pages 1114–1120. ELRA, 2012.

[17] D. Talkin. A Robust Algorithm for Pitch Tracking. In W. B. Kleijn and K. K. Paliwal, editors, *Speech coding and synthesis*, pages 495–517. Elsevier, 1995.

[18] N. Weibel, A. Fouse, E. Hutchins, and J. D. Hollan. Supporting an Integrated Paper-Digital Workflow for Observational Research. In *Proceedings of IUI 2011, International Conference on Intelligent User Interfaces*, Palo Alto, CA, USA, February 2011.

[19] D. Wigboldus, J. Sherman, H. Franzese, and A. Knippenberg. Capacity and comprehension: Spontaneous stereotyping under cognitive load. *Social Cognition*, 22(3):292–309, 2004.