# Estimation of Presentations Skills Based on Slides and Audio Features

Gonzalo Luzardo, Bruno Guamán, Katherine Chiluiza,
Jaime Castells and Xavier Ochoa
Escuela Superior Politécnica del Litoral
Guayaquil, Guayas, Ecuador
gluzardo@cti.espol.edu.ec, bguaman@cti.espol.edu.ec, kchilui@cti.espol.edu.ec,
jaime.castells@cti.espol.edu.ec, xavier@cti.espol.edu.ec

## ABSTRACT

This paper proposes a simple estimation of the quality of student oral presentations. It is based on the study and analysis of features extracted from the audio and digital slides of 448 presentations. The main goal of this work is to automatically predict the values assigned by professors to different criteria in a presentation evaluation rubric. Machine Learning methods were used to create several models that classify students in two clusters: high and low performers. The models created from slide features were accurate up to 65%. The most relevant features for the slide-base models were: number of words, images, and tables, and the maximum font size. The audio-based models reached up to 69% of accuracy, with pitch and filled pauses related features being the most significant. The relatively high degrees of accuracy obtained with these very simple features encourage the development of automatic estimation tools for improving presentation skills.

## Categories and Subject Descriptors

I.5.2 [**Pattern Recognition**]: Design Methodology—*Feature evaluation and selection, Pattern analysis*

## General Terms

Human Factors, Measurement

## Keywords

Multimodal Learning Analytics; presentation skills; slides features; audio features

## 1. INTRODUCTION

The ability to perform a good oral presentation is one of the student outcomes most undergraduate programs aim to develop in their students [30]. There are some courses included in program curricula that encourage the development of such skills. Nevertheless and despite the importance of cultivating these skills, teachers struggle between the time they need to lecture and the one devoted to giving feedback to their students [26] when they do presentations. Lecturers invest much time in attending to these presentations, writing critiques and remarks about students' speech performance and revising digital slides, the common multimedia support for presentations [9]. Giving on-time feedback and nurture presentation skills is not an easy task, some studies focus on the design and development of an instructional approach to improve the student presentation skills and the role of some instructional variables in its development [18] [6] [22].

To optimize professors' time, automatic methods for evaluating and giving feedback could be implemented. In oral presentations, beyond the presenters' skill to transmit their ideas through verbal and non-verbal cues [5], it is well known the importance of using slides as a visual supporting material [16]. The objective of using slides is making presentations more structured and interesting to the audience [9]. Traditionally, automatic evaluation of presentations focus mainly on extraction of just one kind of cues, being those usually visual or audio cues [10]. This work explores several simple audio features based on the prosody of the speech for automatic assessment of the quality of presentations. However, professors use more than just the audio and the video to assess presentations. Another source of evaluation evidence is the quality of the multimedia support materials that are used alongside the presentation, most commonly some type of slide show. Although it has been demonstrated that the slide quality has a direct influence in the perceived quality of a presentation by the audience [8], features related with slides design have received very little research attention. This work will extract features from digital slide files such as readability, number of images, and visual impact among others. This work will combine these two sources into a multimodal analysis of the presentations and estimation of its quality. This analysis will be conducted over an oral presentation dataset that includes 448 individual audio files and the corresponding 86 digital slide files.

The study is structured as follows: first, a section about work related to the topic of this paper is presented; next, the dataset and characteristics of the features in the dataset are included. Section 4, depicts the analysis and results obtained using Machine Learning Methods. In Section 5, the authors discuss the results from previous section. Finally, in Section 6 the conclusions of this paper are given as well as possibilities for further research.


*MLA'14,* November 12, 2014, Istanbul, Turkey.

http://dx.doi.org/10.1145/2666633.2666639.

## 2. RELATED WORK

Teachers need to effectively give feedback to their students. At the same time, students should know beforehand the criteria that are used to evaluate their presentations. These needs lead to the creation of many rubrics for evaluating oral presentations. Some include performance criteria such as: presenters' voice, volume or tone, corporal features and slides' quality. Although, the measurement of these features is usually conducted by other human beings (professors or fellow students) [26], it could also be performed by an automated system based on digital image processing, digital audio processing and machine learning algorithms that could produce the values required in the rubric.

In this work, the student audio during the presentation was recorded for analysis. Several other works report about automated audio features that could be used to determine the quality of the presentation. For example in [28], automatic extracted prosodic features and personality assessments were combined and analyzed to classify speakers as professionals or non-professionals. In another example, the audio has been extracted to analyze how much cognitive work the students do [17]. By analyzing features such as articulation rates and pauses, it could be determined how well the students perform. Also, there has been previous studies in which the audio was extracted from speech to determine how well the person can speak certain language [27] or to determine the liveliness in a presentation [19]. However, to the knowledge of the authors, there exists no literature about how other audio features, such as pitch, could determine the quality of the presentation.

Automatic slides evaluation is an under researched topic. Few studies related to slides characteristics have been developed. An example is PPSGen [20], which generates well-structured presentation slides from academic papers. Due to the fact that academic papers have always a similar structure, this system can produce an acceptable result. Cooper [13] developed a presentation-video-retrieval system using automatically recovered slides and spoken text. In this work, the slides are analyzed from the video recorded presentations. Text is captured from the image sequences via optical character recognition (OCR) and spoken information is also retrieved from video presentations via automatic speech recognition (ASR). This set of features are used to make corrections of indexes and consequently facilitates video retrieval. In addition to the above studies, there are a variety of studies that try to determine which features should be taken into consideration to assess the quality of multimedia support materials. McKenzie [21] presents several guidelines and rubrics related to the quality of slides presentations. These guidelines are related to the text and other elements commonly used in presentations. Regarding to text usage, one guideline proposes to use not more than a dozen of words in a single slide. As for the font size, it is advised that it has to be large enough for being legible and with sufficient contrast between background and text of the slide. However, to the knowledge of the authors, there is no research that has reported about automatic analysis of slides design features and their impact on the perceived quality.

## 3. DATASET

### 3.1 Description

The data analyzed in this work corresponds to the audio and slides information included in the Oral Presentation Quality dataset available to the participants of the Third International Multimodal Learning Analytics Workshop and Challenges (MLA 2014), which seeks to solve the following questions: a) How multimodal techniques can help the assessment of presentation skills?, and b) how good is a group presentation based on the individual performance (audio, video and posture) and the quality of the slides used?

This dataset was composed by 448 multimodal recordings on 86 oral presentations of undergraduate student groups. Each student group contained an average of four speakers. The dataset also included human-coded information about the quality of the presentation. The human coding was recorded with a rubric that measured: a) speech organization, b) volume and voice quality, c) use of language, d) slides presentation quality, e) body language and f) level of confidence during the presentation. Table 1 shows all evaluation criteria used to assess the quality of the oral presentation. The score goes from 1 (low) to 4 (high). The students of each group were evaluated individually using these metrics. The evaluation of the metrics related to the slides was the same for all group members.

For more information about the dataset, the reader could review its description page[1].

### 3.2 Extracted features

In order to obtain the features that were used to predict the quality of the oral presentations, each mode (audio and slides) was analyzed. This section describes these features and the procedure used to extract them.

#### *3.2.1 Slide features*

Each group in the dataset had created a slide presentation. Those slide files were automatically processed to obtain relevant features. The feature selection was guided by the hypothesis that the less text, colors, and objects the slides had, the better they were. This hypothesis is supported by Bulska [8]. With this consideration, two main approaches were followed to obtain the slide features.

The first approach consisted in the individual analysis of the 86 files in PowerPoint format that each group of students presented. This analysis was performed to obtain a first set of features related to the number of images and font sizes used in the slides. A macro was programmed to automatically calculate the total number of words, charts, tables and images, as well as, the minimum and maximum size of the fonts and maximum number of different font sizes per slide. The extracted features from presentations using the macro were the following:

- *Total number of images (TNI)*
- *Minimum font size (MINFS)*
- *Maximum font size (MAXFS)*
- *Maximum number of different font sizes per slide (MAXDFS)*
- *Total number of words (TNW)*
- *Total number of charts (TNC)*

[1]MLA 2014 Oral Presentation Quality Dataset: http://www.sigmla.org/datasets/

Table 1: *Evaluation criteria used for scoring the student oral presentations*

| **Speech Organization** | **Volume / Voice** | | **Language** | **Slides Presentation** | | | **Body Language** | | **Confidence during the presentation** |
|---|---|---|---|---|---|---|---|---|---|
| Structure and Connection of Ideas | Relevant information with good pronunciation | Adequate voice volume for the audience | Language used in presentation according to audience | Grammar | Readability | Impact of the Visual design of the presentation | Posture and Body Language | Eye Contact | Self Confidence and Enthusiasm |

- *Total number of tables (NT)*

The second approach used for extracting relevant slide features was to analysis each slide of the PowerPoint file as a gray JPEG image to calculate its entropy. The entropy of the image is a proxy to determine the level of contrast in the slides. The objective to calculate the entropy is to identify how readable were the slides in a presentation. For example, it is common in undergraduate student presentations to select background colors similar to the font color in the slide. This selection results in slides that are difficult to be read by the audience. The entropy of an image, considering the spread of grey level values in its histogram, measures how much different tonalities are used. Thus, a flat image will have a zero entropy. On the other hand, a high entropy value is obtained when pixels take values all over the available range [23]. The entropy of each slide was calculated and then its maximum, minimum, average and standard deviation were computed too. A MatLab script was implemented to analyze each PowerPoint file and extract the following features:

- *Minimum Entropy value (MINENT)*
- *Maximum Entropy value (MAXENT)*
- *Average of Entropy values(AVGENT)*
- *Std. dev. of Entropy values (STDENT)*

### 3.2.2 Audio features

In order to avoid over fitting, noisy audio files were removed from the dataset, resulting in 384 audio recordings to be processed. Three different approaches for extracting audio features were used.

The first approach was an analysis of some prosodic features which have been shown to characterize the liveliness of the speaker. A lively voice is described as one that varies in intonation, rhythm and loudness; qualities that can be obtained analyzing the pitch of the speaker [19]. The software Speech Analyzer v3.1[2] was used to get the pitch using a 20ms windows size. This size value was selected since it corresponds to a maximum pitch period. This information was used to calculate the following features for each student intervention:

- *Minimum pitch value (MINP)*
- *Maximum pitch value (MAXP)*
- *Average pitch value (AVGP)*
- *Pitch standard deviation value (STDP)*

[2]Speech Analyzer Software
http://www.sil.org/computing/sa/index.htm

In order to improve pitch extraction a gender discrimination analysis was performed. Each audio file was tagged to specify the gender of the speaker. A Snack Sound Toolkit [3] script was implemented to extract the pitch from the audio using ESPS pitch tracker method [31]. Pitch extraction was performed using a 75ms window size and 60-400Hz for males and 75-600Hz for females pitch window size. The following features were extracted using this process:

- *Minimum pitch value (MINESPSP)*
- *Maximum pitch value (MAXESPSP)*
- *Average pitch value (AVGESPSP)*
- *Pitch standard deviation value (STDESPSP)*

The second approach was a speech rate analysis, which has demonstrated correlation with important aspects of a speaker such as persuasiveness and credibility [4]. This skills allow the speakers to transmit their ideas with more confidence. The speech rate related features were extracted and are described as follows:

- *Speech rate (SR)*: The number of syllables divided by the total duration in seconds of each participant presentation.
- *Articulation rate (AR)*: The number of syllables divided by the speaking time.
- *The Average Syllable duration (ASD)*: The ratio of the speaking time over the number of syllables.

The number of syllables was extracted by counting the detected syllables nuclei. The SyllableNuclei Praat script by Nivja de Jong was used for the extraction of number of syllables and thus obtaining the speech rate features mentioned above. A syllable nuclei is identified by locating the peak of a syllable which is usually present in the vowel of the syllable. For detecting each syllable nuclei the intensity peaks were calculated along all the audio file. Since intensity peaks that are preceded by dips are considered to be potential syllable nuclei, only those peaks with a defined dip value were kept, the rest discarded. Also, since pitch extraction is very effective for separating the audio in silence and voiced parts, it was calculated throughout all the presentation audio file, aiming to discard peaks that are present during silence segments of the audio. The rest of peaks were considered as syllable nuclei. A more detailed explanation of how the script works, together with a validation of its effectiveness can be found in [15].

[3]Snack Sound Toolkit
http://www.speech.kth.se/snack/

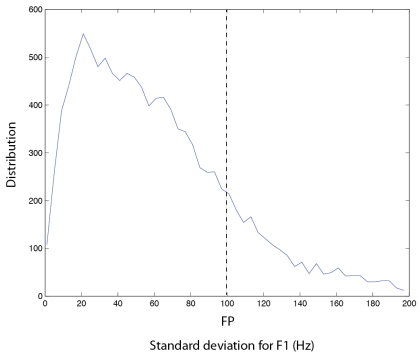


Figure 1: Distribution of F1 standard deviations.

As to ASD, Koopmans-van Beinum and Van Donzel [24] also found that greater values of ASD occurs when the speaker pauses for structuring ideas or making personal comments, making more natural and comprehensible a speech. It is expected that excessively lower and greater values correlates negatively with the perceived speech quality.

The third approach was a speech fluency analysis, based on measuring the presence of filled pauses due to their relation with the fluency of a speech [2]. Since formant information of the speech outstands in the detection of filled pauses, a formant based technique similar to the presented by Audhkhasi [3] was used.

Wavesurfer[4] was also used to compute the first and second formants F1 and F2, which were calculated at a frame rate of 10ms. The standard deviation of each formant F1 and F2, measured as STDF1 and STDF2 respectively, was computed separately to determine the stability of the formants with a window of 6 frames. STDF1 and STDF2 were distributed on histograms of 50 bins having values from 0 to 200 Hz. Fig. 1 shows a histogram calculated for an audio file from the dataset.

The distribution of STDF1 and STDF2 for normal speech tends to be scattered over most of the histogram, in contrast to the speech with filled pauses, in which the distribution is accumulated at the left side in a given frequency point (FP); this is because formants tend to keep static when a filled pause occurs. As a measurement of the filled pause presence, the ratio of the sum of the frequencies at the left side of a given frequency point in the histogram was divided over the sum of all its values for F1 and F2; these formant features are named F1R and F2R when FP = 100Hz and F1R2 and F2R2 when FP = 40Hz. Only the left side of the histogram was considered for the ratio, because as mentioned before, the greater number of filled pauses presence, the greater the accumulation of lower standard deviation of frequencies in the histogram. Thus, it is expected that the participants with greater number of filled pauses in their presentations have greater F1R, F2R, F1R2 and F2R2 values.

The formant features are described as follows:

- *F1R, F1R2, F2R, F2R2*: The accumulation of the left-most side histogram of standard deviation of the frequencies of formants F1 (with FP = 100Hz and FP = 40Hz) and F2 (with FP = 100 and FP = 40Hz), divided by the sum of all the histogram frequencies for each formant.

[4]Wavesurfer Software
http://sourceforge.net/projects/wavesurfer/

Table 2: *Evaluation criteria used for predicting levels of domain related to presentation skills*

| **For slide features** |
|---|
| Readability (RD) |
| Impact of the Visual design of the presentation (IVD) |
| **For audio features** |
| Relevant information with good pronunciation (RIGP) |
| Adequate voice volume for the audience (AVV) |
| Self Confidence and Enthusiasm (SCE) |

# 4. ANALYSIS AND RESULTS

## 4.1 Analysis Methodology

Two evaluation approaches were conducted to predict the presentation skills of the students. The first approach was performed using the features of the slides and the second using the extracted audio features. Each approach was paired with the corresponding set of criteria (coded by humans) which relate to the features being measured. Those criteria are presented in Table 2.

Each evaluation was composed of three phases: preprocessing, training and testing. The preprocessing phase refers to the action of filtering the dataset in order to improve the classification results. In the preprocessing phase all features were normalized and centered. Then a correlation ranking metric was used to perform a feature selection. The second phase was training. In this phase, the features of a random sample of 90% of the students are fed into different machine learning algorithms in order to create a classification model. Finally, the features of the remaining 10% of the students are used as a test set to validate the accuracy and performance of the classification model. The training and testing are repeated 10 times in what it is called a 10-fold cross-validation. The software Weka v3.6.11[5] was used in all phases.

The classification of students were discretized into high and low performers, due to giving better classification accuracy results than using a wide range of values. Thereby, the human generated criteria were combined and clustered into two categories before the training phase. For example, to determine the quality of slide presentation, the criteria "Readability" (RD) and the criteria "Impact of the visual design of the presentation" (IVD) are added. This new criteria (RD+IVD) has values from 2 to 8 because the individual criteria had a value between 1 and 4. A simple rule is used to group those values into two different categories. If the RD+IVD value is lower than 5, those students belong to the first class (C1). On the other hand, if the RD+IVD value of the student is higher or equal to 5, it is considered in the second class (C2). This step was performed both to the slide and audio criteria. Tables 3 and 4 show the classes created.

In the training and testing phases several basic supervised classifiers were used. Two decision tree classifiers: J48 [29] and Random Forest [7]; one rule-based classifier: JRIP [12];

[5]Weka Software
http://www.cs.waikato.ac.nz/ml/weka/

Table 3: *Classes for slides classification*

| Human-codes | Class | Range |
|---|---|---|
| RD+IVD | C1 | >= 0 and < 5 |
| | C2 | >= 5 and <= 8 |

Table 4: *Classes for audio classification*

| Human-codes used | Class | Range |
|---|---|---|
| RIGP+AVV+SCE | C1 | >= 0 and <= 2.5 |
| | C2 | > 2.5 and <= 4 |
| RIGP | C1 | >= 0 and <= 2.5 |
| | C2 | > 2.5 and <= 4 |
| AVV | C1 | >= 0 and <= 2.5 |
| | C2 | > 2.5 <= 4 |
| SCE | C1 | >= 0 and <= 2.5 |
| | C2 | > 2.5 and <= 4 |

two lazy classifiers: IBk (k=5) [1] and KStar [11]; and two function classifiers were Logistic Regression [25] and Support Vector Machine (SVM) [14]. This phase was performed using the default values of the Weka explorer.

Finally, a classifier feature evaluation, using the classifier with best results, was performed to analyze which features it considered the more relevant.

## 4.2 Results

### 4.2.1 Slides evaluation approach

In preprocessing, features Total Number of Workds (*TNW*), Total Number of Tables (*NT*), Total number of images (*TNI*) and Maximum Font Size (*MAXFS*) had a correlation value greater or equal than 0.1 and were selected for the model. Then, the training and testing phases were performed; Table 5 shows the results. While the KStar and Random Forest classifiers have better results than other classifiers, it can be concluded that a 65% of accuracy can be reached. This means that the selected features can determine if a student will obtain a low or high performance on the combination of "Readability" and "Impact of Visual Design on the Presentation" criteria (RD+IVD) 65% of the time.

The KStar was selected to perform the classifier feature evaluation. Table 6 shows the ranked values obtained.

### 4.2.2 Audio evaluation approach

Table 5: *Obtained results for the classifiers using slides features*

| Classifier | accuracy | F-measure | ROC Area |
|---|---|---|---|
| J48 | 0.588 | 0.577 | 0.572 |
| **Random Forest** | **0.671** | **0.671** | **0.699** |
| IBk | 0.612 | 0.605 | 0.680 |
| **KStar** | **0.694** | **0.691** | **0.697** |
| JRip | 0.647 | 0.644 | 0.630 |
| Logistic | 0.624 | 0.612 | 0.685 |
| SVM | 0.565 | 0.499 | 0.519 |

Table 6: *Slides features ranked by KStar*

| Rank | Feature | Ranked value |
|---|---|---|
| 1 | TNW | 64.706 |
| 2 | TNI | 61.176 |
| 3 | NT | 58.824 |
| 4 | MAXFS | 58.824 |

In this approach, four evaluations were conducted, one for each combined criteria presented in Table 4. For the "Relevant information with good pronunciation" (*RIGP*) and "Adequate voice volume for the audience" (*AVV*) evaluations, an additional preprocessing step was carried out. First, in both datasets, students with values lower than 4 and greater than 2 were discarded, since they produced high level of unbalance. Later, the resulting students were separated in two folds (C1 and C2) balanced by using Spread Subsample filter from Weka. The results are presented in Table 7.

For the classification on the combined criteria *RIGP + AVV + SCE*, models were able to reach a 60% of precision, with J48 and Random Forest performing better than other classifiers. The best ranked features are shown in Table 8. Pitch standard deviation (*STDESPSP*) and average (*AVGESPSP*) are the two most important features.

For the classification based on the performance for the individual criteria of "Relevant information with good pronunciation" (*RIGP*), "Adequate voice volume for the audience" (*AVV*) and "Self Confidence and Enthusiasm" (*SCE*) the results are also shown on Table 7. The best classification reached 67% for *RIGP* (Logistic classifier), 69% for *AVV* (JRIP classifier) and 63% for *SCE* (Logistic classifier).

The best classifying features extracted for the pronunciation (*RIGP*) are presented in Table 9. Pitch and filled pauses features seem to be important to predict the quality of the pronunciation. For the volume (*AVV*) the results can be seen in Table 10. Here pitch related features are the most important to obtain high scores related to volume and voice quality evaluation. Finally, for the confidence and enthusiasm (SCE), the results are presented in Table 11. Filled pauses feature and the articulation rate (*AR*) are deemed influential in the perception of the self-confidence and enthusiasm of the student.

# 5. DISCUSSION

Most of the results found in the previous analysis confirm the hypothesis of the authors, but some were unexpected and warrant further research.

## 5.1 Slides evaluation discussion

In the analysis of the slide features, it was expected from existing literature and common sense, that the Total Number of Words (*TNW*) is the most important feature to determine if slide presentation is good or not. Having slides with large amounts of text is one of the main signals that the visual aid is not really helping the presentation. The number of images (*TNI*) is the second most important feature. The impact of this feature is positive, meaning that a large number of images is correlated with better slides. Number of tables (*NT*) is third with a negative relation (more tables, lower grade) and the Maximum Font Size (*MAXFS*) is fourth with a positive relation (larger font, higher grade).

Table 7: *Obtained results for the classifiers using audio features*

| Learner | accuracy | | | | F-measure | | | | ROC Area | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | RIGP+ AVV+ SCE | RIGP | AVV | SCE | RIGP+ AVV+ SCE | RIGP | AVV | SCE | RIGP+ AVV+ SCE | RIGP | AVV | SCE |
| J48 | 0.612 | 0.634 | 0.578 | 0.623 | 0.594 | 0.604 | 0.568 | 0.623 | 0.598 | 0.549 | 0.565 | 0.648 |
| Random Forest | 0.605 | 0.625 | 0.627 | 0.636 | 0.604 | 0.619 | 0.626 | 0.636 | 0.620 | 0.636 | 0.681 | 0.647 |
| IBk | 0.589 | 0.598 | 0.602 | 0.612 | 0.589 | 0.593 | 0.602 | 0.613 | 0.628 | 0.625 | 0.611 | 0.644 |
| Kstar | 0.548 | 0.464 | 0.590 | 0.605 | 0.542 | 0.464 | 0.582 | 0.603 | 0.587 | 0.510 | 0.618 | 0.619 |
| JRIP | 0.594 | 0.643 | 0.687 | 0.597 | 0.593 | 0.641 | 0.686 | 0.594 | 0.593 | 0.610. | 0.685 | 0.580 |
| Logistic | 0.587 | 0.670 | 0.639 | 0.633 | 0.584 | 0.669 | 0.638 | 0.628 | 0.608 | 0.716 | 0.663 | 0.648 |
| SVM | 0.602 | 0.598 | 0.645 | 0.618 | 0.596 | 0.589 | 0.643 | 0.605 | 0.596 | 0.598 | 0.645 | 0.601 |

Table 8: *Audio features based on RIGP + AVV + SCE ranked by J48*

| Rank | Feature | Ranked value |
|---|---|---|
| 1 | STDESPSP | 60.207 |
| 2 | AVGESPSP | 59.948 |
| 3 | F2R | 57.623 |
| 4 | F2R2 | 55.297 |
| 5 | AVGP | 54.78 |
| 6 | MAXP | 52.455 |
| 7 | F1R2 | 50.904 |
| 8 | AR | 50.904 |
| 9 | SR | 50.904 |
| 10 | ASD | 49.871 |

Table 9: *Audio features based on RIGP ranked by Logistic*

| Rank | Feature | Ranked value |
|---|---|---|
| 1 | AVGP | 63.3929 |
| 2 | F2R2 | 62.5 |
| 3 | STDESPSP | 61.6071 |
| 4 | MINESPSP | 59.8214 |
| 5 | F2R | 59.8214 |
| 6 | STDP | 58.9286 |
| 7 | AVGESPSP | 58.9286 |
| 8 | MAXESPSP | 58.0357 |
| 9 | SR | 55.3571 |
| 10 | AR | 53.5714 |
| 11 | ASD | 53.5714 |
| 12 | F1R2 | 51.7857 |

Table 10: *Audio features based on AVV ranked by JRip*

| Rank | Feature | Ranked value |
|---|---|---|
| 1 | STDESPSP | 66.8675 |
| 2 | AVGESPSP | 65.6627 |
| 3 | F2R2 | 63.253 |
| 4 | AVGP | 53.253 |
| 5 | MINP | 61.4458 |
| 6 | ASD | 57.8313 |
| 7 | F2R | 56.6265 |
| 8 | SR | 56.0241 |
| 9 | MAXP | 55.4217 |
| 10 | F1R2 | 54.8193 |
| 11 | STDP | 51.8072 |
| 12 | AR | 51.8072 |
| 13 | MAXESPSP | 43.9759 |

Table 11: *Audio features based on SCE ranked by Random Forest*

| Rank | Feature | Ranked value |
|---|---|---|
| 1 | F2R | 58.39793 |
| 2 | AVGP | 58.13953 |
| 3 | F2R2 | 56.07235 |
| 4 | AR | 54.52196 |
| 5 | STDESPSP | 54.00517 |
| 6 | AVGESPSP | 53.74677 |
| 7 | F1R2 | 51.67959 |
| 8 | ASD | 50.90439 |
| 9 | MINESPSP | 50.64599 |
| 10 | MAXP | 50.64599 |
| 11 | F1R | 50.3876 |
| 12 | SR | 47.54522 |

Other slide features that were thought to be important resulted in having no relevance for the analysis. The entropy obtained from the slides, though appearing promising, was found to be not important to determine the quality of the slides. For example, the standard deviation of the entropy (*STDENT*) feature cannot predict the readability of a presentation because, generally, the design of the slides do not vary throughout the presentation. The minimum entropy (*MINENT*) feature was not relevant either since most presentations contained slides which had a single title, and a plain background, which affected this feature and it was not relevant for the whole presentation. The feature that was expected to be of high significance, the average entropy (*AVGENT*) feature decreased the classifier accuracy value in all cases.

A further analysis could be conducted by measuring the time of the students per slide, since the some slides have different purposes compared to others. In future works, this could help to obtain new features, such as number of slides words, images, or charts, per time period. This could be relevant as well to decide if a presentation was good or not.

## 5.2 Audio evaluation discussion

As for audio analysis, in all the criteria, features related to filled pauses were significant into the estimation of the performance. In general, it was found that F2 related features were more relevant than F1 ones, since the F2 has been shown to be more precise when detecting filled pauses [3]. Moreover, pronunciation (*RIGP*) and confidence and enthusiasm (*SCE*) are highly related to these features. This result

reinforces the common sense notion that less use of filled pauses is an indicator of a good speaker.

Furthermore, for pitch related features, it is noted that in most evaluations the minimum and maximum pitch values are not good indicators for detecting presentation skills, in contrast to standard deviation, which has been found to be highly correlated to the perceived speech quality, since it measures the variation of intonation of the speaker that is one of the criteria for a successful presentation [10]. However, the positive significance of the average pitch in the pronunciation, volume, and confidence and enthusiasm criteria is not easy to explain. One interpretation is that gender plays a role in the perceived quality of the presentation, given that pitch is also an indicator of gender. While not conclusive, this result warrants further research in the topic.

Regarding speech rate features, it has been found that they have relevance only when it comes to evaluation of self-confidence in a speech, which is a desired skill for evaluation. In this study, only the articulation rate ($AR$) was significant, in contrast with speech rate ($SR$); this result can be explained by the existence of long pauses during the speech for presenting multimedia material or due to technical issues during the presentation. In this case, $SR$ may have been affected since it considers the length of all the audio file, contrary to $AR$, which only considers the duration of the speaking time giving a more precise measurement of speech rate.

# 6. CONCLUSIONS

This paper presents the estimation of quality of student presentations, measured using human-generated criteria, through models created from features extracted from slides and recorded audio of those presentations.

The features extracted from slides, such as number of words, number of tables, and number of images were able to support a model that reached 65% of accuracy classifying between what the professor considered good and bad slide presentations. The generated model could be used to construct a slide analysis tool that could provide automatic feedback to the students before the presentation, only requiring them to upload the slides. The proposed tool could provide direct information on what aspects of the slide presentation to improve, for example: "Reduce the amount of words per slides", "Use bigger fonts", etc.

The features extracted from the audio recordings, related to use of filled pauses and the pitch average and variation, were able to produce models that classified between good and bad oral presentations according to different criteria, such as pronunciation, volume and enthusiasm between 60% and 67% of the time. While the filled pauses features could be used to build tools similar to the one proposed for the slides, the relation of pitch with gender warrants further research, in the role that being male or female has on the perceived presentation quality, in order to discard any experimental bias.

A final conclusion of the work is that even simple features from audio and the slide files could already produce better-than-guessing models to estimate the grade that a human would assign to different criteria of a student presentation rubric. These results are encouraging to continue the research on other multimodal features that combined with the simple features proposed in this work, could improve the accuracy of automatic estimation tools. These tools can be used both by the teacher in providing feedback to the students, or even by the students to receive early feedback for their presentation skills before performing in front of professors or fellow students.

# 7. ACKNOWLEDGMENTS

The authors want to acknowledge the support of the VLIR-UOS project ZEIN2010RIP09 and the SENESCYT Project "Andamios".

# 8. REFERENCES

[1] D. Aha and D. Kibler. Instance-based learning algorithms. *Machine Learning*, 6:37–66, 1991.

[2] G. An, D.-G. Brizan, and A. Rosenberg. Detecting laughter and filled pauses using syllable-based features. In *INTERSPEECH*, pages 178–181, 2013.

[3] K. Audhkhasi, K. Kandhway, O. D. Deshmukh, and A. Verma. Formant-based technique for automatic filled-pause detection in spontaneous spoken english. In *Acoustics, Speech and Signal Processing, 2009. ICASSP 2009. IEEE International Conference on*, pages 4857–4860. IEEE, 2009.

[4] J. Benkí, J. Broome, F. Conrad, R. Groves, and F. Kreuter. Effects of speech rate, pitch, and pausing on survey participation decisions. In *Proc. Surv. Res. Meth. Sect. Am. Statist. Ass*, 2011.

[5] E. Bhattacharyya and H. Bt Idrus. To speak like an engineer: Communicative competence in technical oral presentations through the lens of students and industry practitioners. In *Teaching, Assessment and Learning for Engineering (TALE), 2013 IEEE International Conference on*, pages 796–799, Aug 2013.

[6] J. Bourhis and M. Allen. The role of videotaped feedback in the instruction of public speaking: A quantitative synthesis of published empirical research. *Communication Research Reports*, 15(3):256–261, 1998.

[7] L. Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[8] E. Bulska. Good oral presentation of scientific work. *Analytical and bioanalytical chemistry*, 385(3):403–405, 2006.

[9] M. Carter. The use of slides in oral presentations. In *Designing Science Presentations*, pages 191 – 201. Academic Press, San Diego, 2013.

[10] M. Cavanagh, M. Bower, R. Moloney, and N. Sweller. The effect over time of a video-based reflection system on preservice teachers' oral presentations. *Australian Journal of Teacher Education*, 39(6):1, 2014.

[11] J. G. Cleary and L. E. Trigg. K*: An instance-based learner using an entropic distance measure. In *12th International Conference on Machine Learning*, pages 108–114, 1995.

[12] W. W. Cohen. Fast effective rule induction. In *Twelfth International Conference on Machine Learning*, pages 115–123. Morgan Kaufmann, 1995.

[13] M. Cooper. Presentation video retrieval using automatically recovered slide and spoken text, 2013.

[14] N. Cristianini and J. Shawe-Taylor. *An introduction to support vector machines and other kernel-based learning methods*. Cambridge university press, 2000.

[15] N. de Jong and T. Wempe. Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, 41(2):385–390, 2009.

[16] N. Erdemir. The effect of powerpoint and traditional lectures on students' achievement in physics. *Journal of Turkish Science Education (TUSED)*, 10(3):176, 2011.

[17] K. Gorovoy, J. Tung, and P. Poupart. Automatic speech feature extraction for cognitive load classification. In *Conference of the Canadian Medical and Biological Engineering Society (CMBEC)*, 2010.

[18] L. D. Grez, M. Valcke, and I. Roozen. The impact of an innovative instructional intervention on the acquisition of oral presentation skills in higher education. *Computers & Education*, 53(1):112–120, 2009.

[19] R. Hincks. Processing the prosody of oral presentations. In *InSTIL/ICALL Symposium 2004*, 2004.

[20] Y. Hu and X. Wan. Ppsgen: learning to generate presentation slides for academic papers. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 2099–2105. AAAI Press, 2013.

[21] M. Jamie. Scoring power points. *The Educational Technology Journal*, 10(1), September 2000.

[22] K. K. Jensen and V. Harris. The public speaking portfolio. *Communication Education*, 48(3):211–227, 1999.

[23] A. Khellaf, A. Beghdadi, and H. Dupoisot. Entropic contrast enhancement. *Medical Imaging, IEEE Transactions on*, 10(4):589–592, 1991.

[24] F. J. Koopmans-van Beinum and M. E. van Donzel. Discourse structure and its influence on local speech rate. In *Proceedings of the Institute of Phonetic Sciences*, volume 20, pages 1–11, 1996.

[25] S. le Cessie and J. van Houwelingen. Ridge estimators in logistic regression. *Applied Statistics*, 41(1):191–201, 1992.

[26] D. Magin and P. Helmore. Peer and teacher assessments of oral presentation skills: how reliable are they? *Studies in Higher Education*, 26(3):287–298, 2001.

[27] Y. Ming, Q. Ruan, and X. Li. Automatic assessment of oral mandarin proficiency based on speech recognition and evaluation. In *Educational and Information Technology (ICEIT), 2010 International Conference on*, volume 3, pages V3–37. IEEE, 2010.

[28] G. Mohammadi and A. Vinciarelli. Humans as feature extractors: Combining prosody and personality perception for improved speaking style recognition. In *Systems, Man, and Cybernetics (SMC), 2011 IEEE International Conference on*, pages 363–366, 2011.

[29] J. R. Quinlan. *C4. 5: programs for machine learning*, volume 1. Morgan kaufmann, 1993.

[30] V. N. Shaw. Reading, presentation, and writing skills in content courses. *College Teaching*, 47(4):153–157, 1999.

[31] D. Talkin. A robust algorithm for pitch tracking (rapt). *Speech coding and synthesis*, 495:518, 1995.