Acoustic and Linguistic Analysis of Interpersonal Communication in Teams

by

Shrivatsa Mishra

B.Tech., Indraprastha Institute of Information Technology Delhi, 2023

M.S., University of Colorado, Boulder, 2025

A thesis submitted to the

Faculty of the Graduate School of the

University of Colorado in partial fulfillment

of the requirements for the degree of

Master of Science in Computer Science

College of Engineering Applied Science

2025

Committee Members:

Theodora Chaspari, Chair

Dr. Peter Foltz

Dr. Maria Pacheco

Mishra, Shrivatsa (MS CSEN)

Acoustic and Linguistic Analysis of Interpersonal Communication in Teams

Thesis directed by Prof. Theodora Chaspari

This thesis examines acoustic and linguistic patterns of teams in interprofessional communication settings, including the cardiac operation theatre and collaborative learning in Science, Technology, Engineering, and Mathematics (STEM). Using speech and language processing techniques, it analyzes communication behaviors to identify factors that impact teamwork.

The first study explores "good" and "poor" communication scenarios in cardiac surgery, focusing on voice prosody and linguistic features. Significant differences were found in acoustic features like frequency and loudness, as well as linguistic aspects such as authenticity and emotional tone between the two communication types.

The second study investigates the impact of gender ratios on team dynamics in online STEM group study sessions. Results indicate significant differences in self-reported measures, linguistic patterns (i.e., analytical thinking, clout, speaking rate, tone), and acoustic features (i.e., pitch, loudness, and jitter) based on team composition.

This thesis highlights the potential of speech to be used as an unobtrusive marker for improving communication and collaboration in teams and could potentially inform training programs to enhance team functioning.

Contents

Chapter

Tables

Table

Figures

Figure

Chapter 1

Acoustic and Linguistic Analysis in Cardiac Surgery

Roughly 350,000 to 500,000 patients undergo cardiac surgery each year, of these 28,000 will have an adverse event, with one third of deaths associated with coronary artery bypass graft (CABG) operations likely being preventable [17]. Many of these errors stem from miscommunication caused by unclear language, omissions, assumptions, or distractions, leading to confusion and mistakes [31, 6]. This is exacerbated in cases where the team is unfamiliar with each other [11]. Effective communication among surgical team members is crucial for patient safety, successful outcomes, task coordination, situational awareness, and rapid response to challenges. [8]. Behavioral data analytics leverage multimodal data and artificial intelligence (AI) methods and can offer valuable insights into communication cues between team members during cardiac surgery [21]. By integrating various data sources such as audio, video, and physiology, AI algorithms can capture team members' tone of voice, language patterns, and body language, ultimately distinguishing between positive and negative interprofessional communication behaviors. Speech signals can serve as indicators of both positive and negative interprofessional communication cues [12]. A recent study that conducted analysis of conversation data during team interaction indicates that team performance can be predicted with 91% accuracy based on acoustic features[1], with the addition of linguistic features improving the Mean Squared Error (MSE) by over 10 points [24]. In another study, speech features extracted from conversations between team members were significantly correlated with team collaboration effort [27]. Identifying the vocal behaviors that affect team functioning on a moment-to-moment basis can yield unique insights for supporting personalized team

training (e.g., video playback [7]) aimed at proactively addressing communication challenges, optimizing collaboration, and ensuring safer, more efficient cardiac surgery procedures. This paper investigates the extent to which acoustic and linguistic measures extracted from speech during two phases of simulated cardiac surgery operations vary depending on the quality of interprofessional communication behaviors between team members.

1.1 Methods

Study Design and Setting: Data for this study originated from four scripted simulation videos that were produced for educational purposes on cardiac OR team training by the American Society of Extracorporeal Technology (AmSECT). The videos used in this study were produced in 2013 by the American Society of Extracorporeal Technology (AmSECT) and the International Consortium for Evidence-Based Perfusion (ICEBP) as part of the Cardiac Surgical Team Training Video Project. The scenarios scripts were created by a multidisciplinary group composed by cardiac surgery experts (surgeons, anesthesiologists and perfusionists) and human factors/medical simulation experts to demonstrate specific types of behaviors in the operating room. Data includes one scripted cardiac scripted surgery video displaying "poor" interprofessional communication and one video of a "good" communication example. Each simulation covers two phases: the pre-operative briefing and the cardiopulmonary bypass (CBP) phase.

Population: Participants of the simulations were four selected representative primary cardiac team members, including the attending cardiac surgeon, attending anesthesiologist, primary perfusionist, and scrub nurse, playing scripted simulation scenarios.

Procedures: All simulated sessions were video recorded. The video editing took place immediately after the simulations are recorded. Audio was further extracted from the videos and the transcripts were manually examined to ensure the correct start and end time of each dialog turn.

Acoustic Analysis: Acoustic measures indicative of voice prosody and intonation were extracted using the Geneva Minimalistic Acoustic Parameter Set (GeMAPS) [14]. Our analysis includes the mean and standard deviation of Fundamental frequency (F0), Loudness, the first four Mel-

Frequency Cepstral Coefficients (MFCC), Alpha Ratio, Hammarberg Index, Spectral Flux, Jitter and Shimmer, since these measures are theoretically stipulated and empirically validated in association with emotional arousal [14]. Additionally, we examine the falling and rising slopes of F0, as these features have been shown to significantly enhance emotion classification accuracy [5]. Each acoustic measure was extracted over each speaker turn at a short-time scale of 20-60ms, following standard practices in speech analysis for capturing human affect [18].

Linguistic Analysis: Linguistic measures indicative of analytical thinking, clout, authenticity, emotional tone, and positive/negative emotions were extracted from the transcripts through LIWC2020 [29]. These measures were selected due to their association with cognitive and emotional processes occurring in team functioning [24].

We utilized the Man Whitney U test to identify significant differences of the features in question in terms of the different phases and settings of the surgery. Linear mixed-effects model (LME) analysis [26] was further applied to identify significant differences in terms of the linguistic and acoustic measures between the "good" communication and the "poor" communication videos. The LME models accounted for the multiple phases j (i.e., briefing, CPB) and speakers s, within each video i and were defined as:

$$Measure_{i,j,s} = \beta_0 + \beta_1 \cdot Quality_{i,j} + \beta_2 \cdot j + u_s + \epsilon_{ijs} \tag{1.1}$$

where the variables are defined as follows:

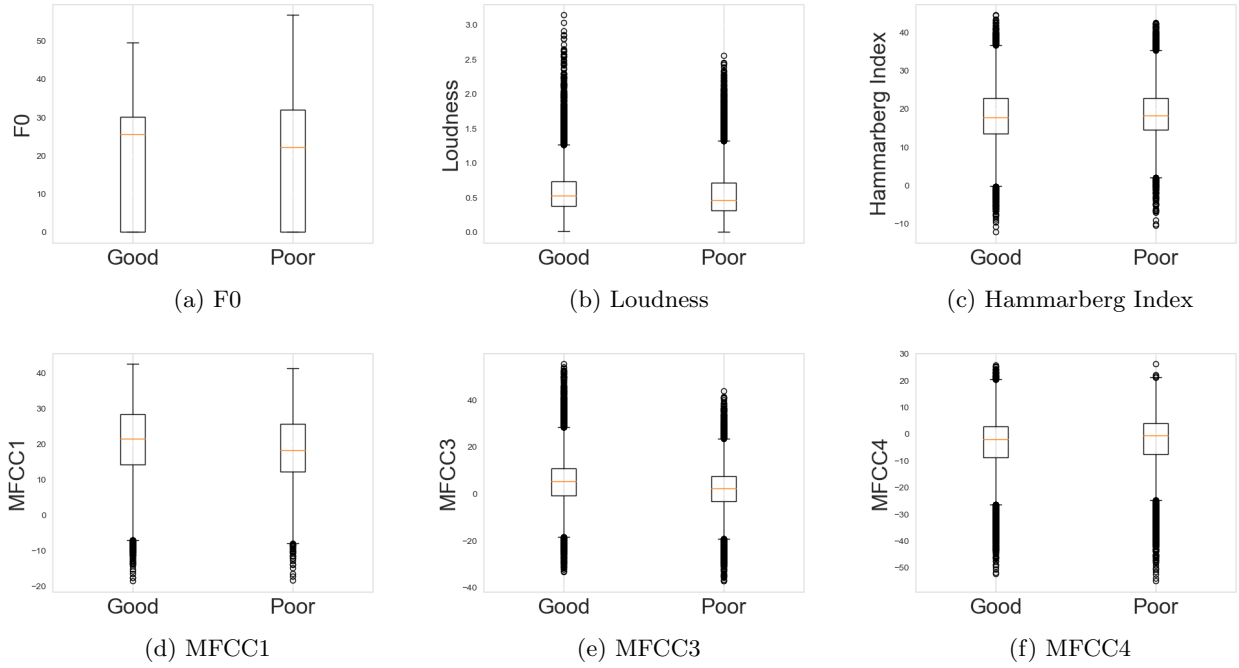
- Measure$_{i,j,s}$ is the linguistic/acoustic measure of speaker s in video i over phase j.
- Quality$_{i,j}$ is the variable representing the communication quality of video i for phase j (i.e., 0 for 'good' and 1 for 'poor').
- j represents the phase of the video (i.e., 0 for briefing, 1 for CPB).
- u_s is the random intercept for speaker s, capturing speaker-specific variability in the measures.
- $\epsilon_{i,j,s}$ is the residual error for the measure of speaker s in video i during task j.

1.2 Results

1.2.1 Acoustic Measures

Acoustic measures are visualized separately for the "good" and "poor" communication videos via box plots, that depict the distribution of each measure per video (Figure 1.1). A preliminary examination reveals a significant difference between the "good" and "poor" communication sessions, particularly in terms of mean fundamental frequency (F0), loudness, and several MFCCs. Specifically, we observe a significant decrease ($p < 0.01$) in F0 in the "Good" session at a median of 27.96 (IQR: 25.22 - 30.42), as compared to the "Poor" session at 30.70 (IQR: 28.16 - 32.84). We also observe a similar trend in loudness being significantly lower ($p < 0.01$) in the "Good" sessions with the median being 0.59 (IQR: 0.50 - 0.75) as compared to the "Poor" sessions at 0.98 (IQR: 0.67 - 1.20), and the Hammarberg Index being significantly lower in the "Good" sessions (Mdn = 14.89 (IQR: 12.71-18.15)), as compared to the "Poor" sessions (Mdn = 16.08 (IQR: 13.76-19.28)). Finally we observe significant difference in terms of the first (Good: Mdn = 22.04 (IQR: 20.21 - 23.96); Poor: Mdn = 24.15 (IQR: 20.95 - 27.45)), third (Good: Mdn = 4.99 (IQR: 2.78 - 8.08); Poor: Mdn = 3.69 (IQR: 0.11 - 6.88)) and fourth (Good: Mdn = -3.50 (IQR: -6.02 - -0.73); Poor: Mdn = -5.96 (IQR: -10.26 - -2.92)) MFCC. These indicate the feasibility of acoustic features in identifying between "good" and "poor" interprofessional communication sessions.

Results from the LME models indicate a significant effect of both the session rating (i.e., good/poor) and phase (i.e., Briefing/CPB) on the acoustic measures on the same measures (Table 1.1). Indicatively, significant differences between the "good" and "poor" surgery sessions are found in terms of the energy/amplitude measures such as loudness (perceived speech intensity). Spectral parameters that overall quantify differences in speech between low and high-frequency regions also depict significant differences between the sessions. Specifically, we observed significant differences between the Briefing and CPB in terms of the alpha ratio (ratio of the summed speech energy from low frequency of 50–1000 Hz and high frequency of 1–5 kHz), Hammarberg index (ratio of the strongest energy peak in the 0-2 kHz region to the strongest peak in the 2-5 kHz region),



(a) F0 (b) Loudness (c) Hammarberg Index

(d) MFCC1 (e) MFCC3 (f) MFCC4

Figure 1.1: Box plots of acoustic measures for "good" and "poor" interpersonal communication sessions.

and spectral flux (temporal difference in spectral content between two consecutive speech frames). Finally, jitter, a frequency-related speech measure that quantifies temporal deviations in speech periodicity, also depicted significant differences between the "good" and "poor" session along with the difference phases.

1.2.2 Linguistic Measures

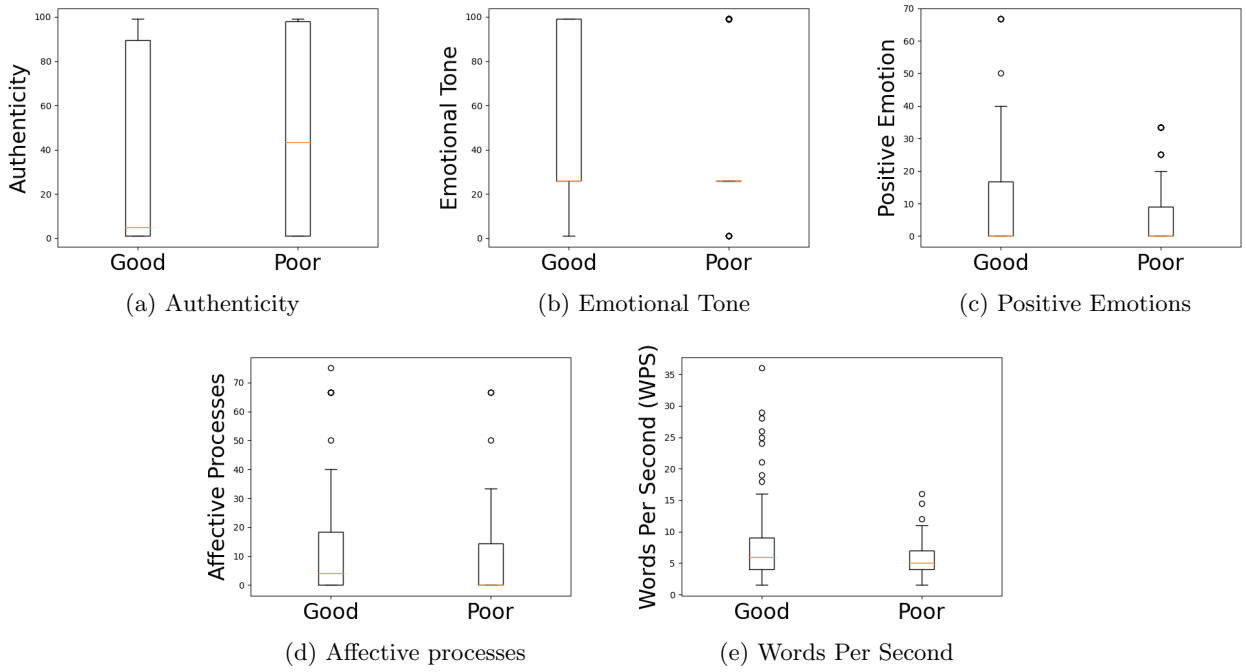
Linguistic measures are also visualized separately for the "good" and "poor" communication videos via box plots, that depict the distribution of each measure per video (Figure 1.2). A preliminary examination suggests significant difference between the "good" and "poor" communication sessions. We observe a significant decrease in words that convey Authenticity for the "Good" session (Mdn = 4.17 (IQR: 1.00-89.63)) as compared to the "Poor" session (Mdn = 43.37 (IQR: 1.00-98.01)). This trend is also present in words conveying emotional tone (Good: Mdn = 4.17 (IQR: 1.00-89.63); Poor: Mdn = 43.37 (IQR: 1.00-98.01)). On the other hand words conveying

Table 1.1: Coefficients and p-values of the linear mixed effects (LME) models comparing the quality of interprofessional communication across surgical phases and communication quality in terms of acoustic features.

Acoustic Measure	**Quality (0: Good/1: Poor)**	**Phase (0: Briefing/1: CPB)**
F0 Mean	$\boldsymbol{\beta_1}$ **= 2.770, p<0.001**	β_2 = -0.674, p = 0.270
F0 Std	$\boldsymbol{\beta_1}$ **= -0.038, p<0.001**	$\boldsymbol{\beta_2}$ **= 0.053, p<0.001**
F0 Rising Slope	β_1 = -4.069, p = 0.852	β_2 = -17.550, p = 0.490
F0 Falling Slope	$\boldsymbol{\beta_1}$ **= -56.452, p<0.001**	$\boldsymbol{\beta_2}$ **= 68.896, p<0.001**
MFCC1	$\boldsymbol{\beta_1}$ **=1.447, p = 0.003**	β_2 = 0.709, p = 0.237
MFCC2	$\boldsymbol{\beta_1}$ **= -2.318, p<0.001**	$\boldsymbol{\beta_2}$ **= 2.965, p<0.001**
MFCC3	$\boldsymbol{\beta_1}$ **= -1.655, p = 0.002**	β_2 = -0.159, p = 0.811
MFCC4	$\boldsymbol{\beta_1}$ **= -3.597, p<0.001**	β_2 = 0.364, p = 0.625
Loudness	$\boldsymbol{\beta_1}$ **= 0.263, p<0.001**	$\boldsymbol{\beta_2}$ **= 0.270, p<0.001**
Alpha Ratio	β_1 = 0.252, p = 0.458	$\boldsymbol{\beta_2}$ **= -2.764, p<0.001**
Hammarberg Index	β_1 = -0.704, p = 0.205	$\boldsymbol{\beta_2}$ **= 3.90, p<0.001**
Spectral Flux	$\boldsymbol{\beta_1}$ **= 0.215, p<0.001**	$\boldsymbol{\beta_2}$ **= 0.235, p = 0.004**
Jitter	$\boldsymbol{\beta_1}$ **= -0.011, p = 0.005**	$\boldsymbol{\beta_2}$ **= 0.018, p<0.001**
Shimmer	β_1 = 0.027, p = 0.276	β_2 = -0.032, p = 0.269

Positive emotions are much more common in the "Good" session (Mdn = 0.00 (IQR: 0.00-20.00)) as compared to the "Poor" session (Mdn = 0.00 (IQR: 0.00-5.55)), as well as words conveying Affective processes (Good: Mdn = 3.78 (IQR: 0.00-20.00); Poor: Mdn = 0.00 (IQR: 0.00-14.29)). There is also a significant increase ($p = 0.04$) in the number of words per second for the "Good" session (Mdn = 5.00 (IQR: 4.00-8.25)), with the "Poor" session having a slower rate of speech (Mdn = 5.00 (IQR: 3.00-6.46))

Results from the LME models indicate a significant effect of the session rating (i.e., good/poor) on the acoustic measures, but little effect of the type of phase (i.e., briefing/CPB) on the same measures (Table 1.2). We observe a significant effect of words related to analytical thinking (i.e., words that point to formal, logical, and hierarchical thinking patterns) when comparing between the Briefing and the CPB. We also observe a significant effect of the quality of interprofessional communication (good/poor) on the number of words conveying Authenticity, Emotional Tone, Positive emotion, and Negative emotion. Finally there is a significant difference in terms speaking rate (i.e., number of words spoken every second) between the "good" and "poor" sessions along with



(a) Authenticity (b) Emotional Tone (c) Positive Emotions

(d) Affective processes (e) Words Per Second

Figure 1.2: Box plots of linguistic measures for "good" and "poor" interpersonal communication sessions.

the difference phases.

1.3 Discussion

The significant differences observed in the acoustic measures between "good" and "poor" cardiac surgery team communication sessions underscore the role of speech acoustics in facilitating effective interpersonal communication. Notably, lower mean fundamental frequency (F0), reduced loudness, and a drop in Hammarberg index characterize "good" communication sessions. These findings suggest that calm, steady, and less intense speech patterns may contribute to a more conducive communication environment. The additional impact of session phase (Briefing vs. CPB) on parameters like the alpha ratio, Hammarberg index, and spectral flux further indicates the dependence of context on speech acoustics. The decrease in alpha ratio from Briefing to CPB along with the increase in spectral flux suggests greater variability in the power spectrum over the phases. Furthermore, the significant differences in jitter–a measure of speech periodicity that

Table 1.2: Coefficients and p-values of the linear mixed effects (LME) models comparing the quality of interprofessional communication across surgical phases and communication quality in terms of linguistic features.

Acoustic Measure	**Quality (0: Good/1: Poor)**	**Phase (0: Briefing/1: CPB)**
Analytical Thinking	β_1 = -7.679, p = 0.079	$\boldsymbol{\beta_2}$ **= 11.693, p = 0.008**
Clout	β_1 = -6.699, p = 0.059	β_2 = -4.477, p = 0.326
Authenticity	$\boldsymbol{\beta_1}$ **= 13.094, p = 0.005**	β_2 = 9.843, p = 0.155
Emotional Tone	$\boldsymbol{\beta_1}$ **= -19.122, p<0.001**	β_2 = 8.703, p = 0.097
Positive Emotion	$\boldsymbol{\beta_1}$ **= -10.913, p<0.001**	β_2 = 4.692, p = 0.105
Negative Emotion	$\boldsymbol{\beta_1}$ **= 1.993, p = 0.007**	β_2 = 1.276, p = 0.163
Words Per Second	$\boldsymbol{\beta_1}$ **= -1.121, p = 0.018**	$\boldsymbol{\beta_2}$ **= -3.634, p<0.001**

increases from the briefing phase to the CPB phase–suggest that vocal stability is a key feature of effective communication, particularly during the high-pressure phases of surgery.

Linguistic analysis revealed that sessions characterized with "good" interprofessional communication depicted higher word rate, a lower proportion of words reflecting authenticity, and higher positive emotional tone, compared to "poor" sessions. These findings suggest that effective communication in these sessions may rely on a combination of fluency and an emphasis on positivity, which together create a more engaging and favorable interaction. The LME model also suggests a positive correlation between emotionally negative language and "poor" sessions, along with a similar negative correlation with emotionally positive language. Words associated with analytical thinking also show an increase during the CPB phase, as compared to the briefing phase, further highlighting the structured and goal-oriented nature of effective communication at the surgical stage. Interestingly, while phases influenced some linguistic patterns, session quality predominantly drove the observed differences. "Good" communication sessions consistently showed a higher frequency of words indicating positivity and emotional balance, reinforcing the notion that collaborative environments are marked by mutual trust and clarity.

Despite more than 20 years have passed since the publication of the seminal U.S. Institute of Medicine's report "To Err is Human" [9], adverse events in surgery remain frequent and preventable [10]. There is urgent need to enhance patient safety using innovative approaches. Our

findings highlight the potential of leveraging quantitative insights from speech analytics in interventions aimed at enhancing team communication in surgical settings. Training programs could focus on promoting speech patterns characterized by calmness, steadiness, and reduced intensity, as these appear to foster better interpersonal dynamics. Additionally, emphasizing the use of positive and emotionally balanced language, alongside maintaining a structured and analytical approach during high-pressure phases like CPB, could improve collaboration and decision-making.

Despite the promising results, the small sample size and reliance on scripted data, highlight the necessity of incorporating more diverse and realistic data from real-life scenarios in subsequent analyses. The limited sample size also hindered the feasibility of automatically detecting sessions with "poor" interprofessional communication. Moreover, the phase-dependent differences suggest that tailored strategies might be needed for different stages of surgical procedures. Additionally, the voice transcription pipeline requires significant improvement to handle the noise generated by surgical tools and the challenges posed by suboptimal microphone placement in real-world settings. This approach will enhance the robustness and generalizability of future findings.

1.4 Conclusion

Our results indicate significant differences in terms of the acoustic and linguistic characteristics of the surgical team members between the "good" and "poor" communication sessions along with between the different phases of the surgery. This suggests that AI systems could rely on such acoustic and linguistic measures to detect subtle nuances in communication dynamics between cardiac surgery team members in-real time, ultimately enabling team alerts and facilitating targeted interventions such as clarifications, conflict resolution, or additional support. Confirmation of these findings in a larger dataset derived from observations in real-life surgery is warranted and ongoing in our own research program.

Chapter 2

Acoustic and Linguistic Analysis in STEM Group Study

Gender equity remains a significant challenge within the fields of science, technology, engineering, and mathematics (STEM), where women are consistently underrepresented and underpaid [25]. In particular, women are known to leave engineering at higher rates than men and generally have higher GPAs while enrolled [2]. Collaborative learning techniques, such as team projects, are often proposed as a solution to this problem [4, 13]. Such collaborative learning curricula have shown some success in improving the retention of male extroverts [15],but not of other groups . In fact, as far as women are concerned, there is some reason to suspect that team projects might actually accelerate attrition: women in engineering and technical disciplines frequently report negative team experiences which may make them question their place in the discipline[32].

The existing literature provides information on general gender differences in team interactions and the influence of campus culture on engineering identity and teamwork [33, 30], but a targeted examination of the effect of varying gender proportions on team performance and perceived success in undergraduate STEM is crucial to inform pedagogical practices and foster more equitable and effective learning environments. This study seeks to address this gap by exploring how different gender ratios in undergraduate STEM teams correlate with team outcomes and students' perceptions of their success.

2.1 Methods

2.1.1 Study Design

The study spanned a total duration of 9 hours and 30 minutes, divided over five days. The study was conducted online via Zoom, where participants engaged in group study sessions. Participants first entered a waiting room and changed their profile ID to the assigned code name for anonymization before joining the main session.

The study consisted of three key events:

2.1.1.1 Enrollment

The enrollment process on Day 1 lasted for a total of 60 minutes, beginning with 15 minutes dedicated to obtaining informed consent from participants. Following this, participants spent 45 minutes completing questionnaires designed to assess stress levels, Big Five personality traits, demographics, and other relevant factors.

2.1.1.2 Group Study Sessions

Each group study session, conducted on five days, lasted for 90 minutes each and included four participants per team. The teams were structured to examine group dynamics, with some consisting of one female and three males, while others had an equal gender distribution of two males and two females. To observe temporal changes in communication behavior, the composition of each team remained consistent across all five sessions. Each session followed a structured format:

- **Mood Scaling (5 minutes)**: Pre-session assessment using the Positive and Negative Affect Schedule (PANAS) and the Perceived Stress Scale (PSS).
- **Group Work (60-75 minutes)**: Collaborative problem-solving and coding exercises, with a research assistant present to oversee interactions.

- **Post-Session Measures (10 minutes)**: Mood and stress reassessment, team cohesion and conflict evaluation, and reporting of any adverse events.

2.1.1.3 Exit Session

The exit session on Day 5 lasted for 60 minutes and consisted of two main components. Participants first spent 30 minutes completing the final set of questionnaires. This was followed by 30 minutes of one-on-one discussions with a research assistant via Zoom breakout rooms, where they reflected on their group experience, feelings of inclusion, future academic plans, and perceived opportunities.

2.1.2 Analysis

Beyond the questionnaires, the audio was further extracted from the videos and the transcripts were manually created to ensure the correct start and end time of each dialog turn. This audio and transcript are used for further linguistic and acoustic analysis.

2.1.2.1 Questionnaires

An initial analysis was conducted on self-reported measures, including PANAS scores, perceived stress, relationship conflict, task conflict, conflict management, team cohesion, perceived performance, team satisfaction, and team viability. These factors were compared between teams with balanced and unbalanced gender ratios to assess potential differences in team dynamics and outcomes.

2.1.2.2 Linguistic Analysts

Linguistic analysis consisted of two parts, VADER based Sentiment analysis and LIWC based analysis. First the Valence Aware Dictionary and sEntiment Reasoner(VADER) a lexicon and rule-based sentiment analysis tool was used to extract the positive and negative sentiment present in the transcript for each individual session. Following this linguistic measures indicative of analytical

thinking, clout, authenticity, emotional tone, and positive/negative emotions were extracted from the transcripts through LIWC2020 [29]. LIWC summary features are based on dictionaries that connect important psychosocial constructs and theories with words, phrases, and other linguistic constructions [3]

2.1.2.3 Acoustic Analysis

Acoustic measures indicative of voice prosody and intonation were extracted using the Geneva Minimalistic Acoustic Parameter Set (GeMAPS) [14]. Our analysis includes the mean and standard deviation of Fundamental frequency (F0), Loudness, the first four Mel-Frequency Cepstral Coefficients (MFCC), Alpha Ratio, Hammarberg Index, Spectral Flux, Jitter and Shimmer, due to these measures being theoretically stipulated and empirically validated in association with emotional arousal [14]. Each acoustic measure was extracted over each speaker turn at a short-time scale of 20-60ms, following standard practices in speech analysis for capturing human affect [18].

We tested the normality of the features and based on that used the Kruskal-Wallis test for non normal data with the eta squared as the effect size, while using a t-test for normal data and reported the Cohen's d for the effect size to identify significant differences of the features in question in terms of the different phases and settings of the surgery. Linear mixed-effects model (LME) analysis [26] was further applied to identify significant differences in terms of the linguistic and acoustic measures between the teams with differing gender ratios. The LME models accounted for the multiple days j and speakers s, for each team i and were defined as:

$$Measure_{i,j,s} = \beta_0 + \beta_1 \cdot Ratio_i + \beta_2 \cdot j + u_s + \epsilon \tag{2.1}$$

where the variables are defined as follows:

- Measure$_{i,j,s}$ is the linguistic/acoustic measure of speaker s in team i on day j.
- Ratio$_i$ is the variable representing the gender ratio for the team in question (0: Lower Female Ratio; 1: Higher female ratio).

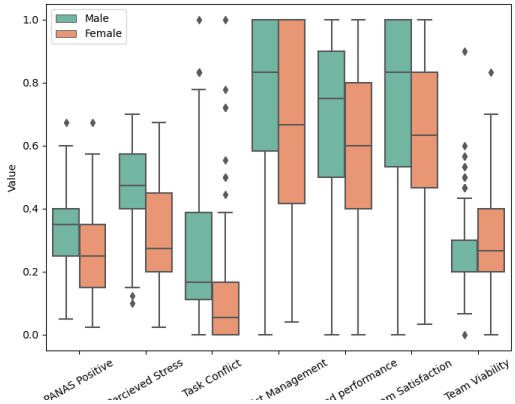
- j represents the day of the task.
- u_s is the random intercept for speaker s, capturing speaker-specific variability in the measures.
- ϵ is the residual error.

2.2 Results

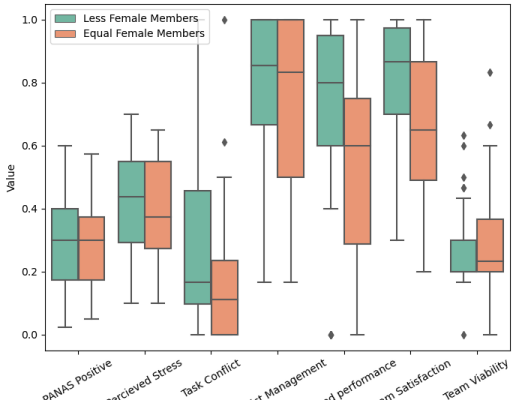
2.2.1 Questionnaire Results

Analysis of the questionnaire results reveals significant differences between male and female participants across multiple measures. These results are summarized in Figure 2.1a. When looking at the questionnaire results we observe a significant difference (p ¡ 0.001; Cohen's d = 0.11) in the reported PANAS Positive score between the male with a higher median score of 0.35 (IQR: 0.25 - 0.40) as compared to female participants with a median of 0.25 (IQR: 0.15 - 0.35). We also observe significantly higher (p<0.001, $\eta^2 = 0.31$) perceived stress for the male participants (Mdn: 0.47 (IQR: 0.40 - 0.57)) as compared to the female participants (Mdn: 0.28 (IQR: 0.20 - 0.45)). Male participants also report significantly higher (p<0.001, $\eta^2 = 0.13$) task conflict at Mdn: 0.17 (IQR: 0.11 - 0.39) as compared to the female participants (Mdn: 0.06 (IQR: 0.00 - 0.17)). This trend is contrasted by a significant difference (p=0.004, $\eta^2 = 0.06$) in the reported Conflict Management being much more for the male participants (Mdn: 0.83 (IQR: 0.58 - 1.00)) as compared to the female participants (Mdn: 0.67 (IQR: 0.42 - 1.00)). Furthermore The perceived performance (p=0.01, $\eta^2 = 0.04$) and the team satisfaction (p<0.001, $\eta^2 = 0.09$) were also both higher for the male participants (Perceived Performance Mdn: 0.75 (IQR: 0.50 - 0.90); Team Satisfaction Mdn: 0.83 (IQR: 0.53 - 1.00))as compared to the female participants (Perceived Performance Mdn: 0.60 (IQR: 0.40 - 0.80); Team Satisfaction Mdn: 0.63 (IQR: 0.47 - 0.83)). Surprisingly the team viability was significantly higher (p=0.005, $\eta^2 = 0.06$) for the female participants (Mdn: 0.27 (IQR: 0.20 - 0.40)) as compared to the male participants (Mdn: 0.20 (IQR: 0.20 - 0.30)).

We also observe the reported Relationship Conflict, Task Conflict, Team Cohesion and the



(a) Comparing Reported Scores for Male and Female Participants



(b) Comparing Reported Scores for Teams with Less and Equal Ratio of Female Members

Figure 2.1: Comparison of reported scores based on gender and team composition.

Team Viability were all significantly better for all members in teams with an equal number of female and male members. However the Perceived Performance, Team satisfaction were significantly better for teams with less female members These results are reported in more detain in Table and Figure 2.1b.

2.2.2 Linguistic Results

A preliminary examination suggests significant difference between the teams. We observe a significant increase ($p<0.001$, $\eta^2 = 0.01$) in words that convey Analytical thinking for teams with less female members (1.00 (IQR: 1.00 - 39.70)) as compared to teams with an equal number of females(Mdn: 1.00 (IQR: 0.00 - 26.10)). A similar trend is observed with words relating clout (Less Female Mdn: 1.00 (IQR: 0.00 - 64.67); Equal Female Mdn: 0.00 (IQR: 0.00 - 4.61); $p<0.001$, $\eta^2 = 0.02$) and authenticity (Less Female Mdn: 15.38 (IQR: 0.00 - 98.19); Equal Female Mdn: 10.18 (IQR: 0.00 - 99.00); $p=0.003$, η^2 ¡ 0.01) with male heavy teams using more than more equally weighted teams. Finally the speed of talking (Words per second) is also significantly faster($p<0.001$, $\eta^2=0.03$) in teams with less female members as compared to the other teams. Results from the LME models also indicate a significant effect of the ratio of female members on the Tone ($p=0.029$)

of the meeting as well as on the neutral sentiments (p=0.026) present in the text.

Table 2.1: Comparison of All Team Member Results Based on Gender Composition of Teams

Feature	**Teams with Less Female Members**		**Teams with Equal Female Members**		P-Value	η^2
	Mean	IQR	Mean	IQR		
Relationship Conflict	0.00	0.00 - 0.25	0.00	0.00 - 0.00	< 0.001	0.21
Task Conflict	0.17	0.10 - 0.46	0.11	0.00 - 0.24	$= 0.004$	0.11
Team Cohesion	0.17	0.00 - 0.70	0.50	0.32 - 0.67	$= 0.03$	0.04
Perceived Performance	0.80	0.60 - 0.95	0.60	0.29 - 0.75	< 0.001	0.31
Team satisfaction	0.87	0.70 - 0.97	0.65	0.49 - 0.87	< 0.001	0.16
Team Viability	0.20	0.20 - 0.30	0.23	0.20 - 0.37	$= 0.001$	0.14

2.2.3 Acoustic Results

A preliminary examination suggests significant difference between the teams, particularly in terms of mean fundamental frequency (F0), loudness, several MFCCs, Jitter, Shimmer, Alpha Ratio and the Hammarberg Index. Specifically, we observe a significant increase ($p<0.001$, $\eta^2=0.07$) in the fundamental frequency for teams with an equal ratio of female members(Mdn: 28.81 (IQR: 26.14 - 33.13)) as compared to the rest(Mdn: 26.38 (IQR: 23.98 - 31.88)). Similarly teams with an equal ratio (Mdn: 0.84 (IQR: 0.57 - 1.10)) also are significantly louder ($p<0.001$, $\eta^2=0.05$) as compared to teams with less number of females (Mdn: 0.70 (IQR: 0.51 - 0.90)). These teams also have a significantly higher MFCC3 (Less Female Mdn: 13.79 (IQR: 8.15 - 19.34); Equal Female Mdn: 19.26 (IQR: 11.83 - 27.03); $p<0.001$, $\eta^2 = 0.12$), while a significantly lower MFCC1 (Less Female Mdn: 27.15 (IQR: 21.40 - 33.07); Equal Female Mdn: 20.28 (IQR: 12.71 - 27.45); $p<0.001$, $\eta^2 = 0.19$), MFCC2 (Less Female Mdn: 9.31 (IQR: 2.55 - 15.58); Equal Female Mdn: 5.97 (IQR: -0.27 - 11.99); $p<0.001$, $\eta^2 = 0.04$) and MFCC4 (Less Female Mdn: 2.25 (IQR: -3.67 - 7.85); Equal Female Mdn: -2.45 (IQR: -9.25 - 3.74); $p<0.001$, $\eta^2 = 0.10$). Alpha Ratio is also significantly higher ($p<0.001$, $\eta^2=0.07$) for teams with an equal number of female members (Mdn: -2.64 (IQR: -8.67 - 3.04)) as compared to less (Mdn: -6.68 (IQR: -11.16 - -0.38)). On the other hand such teams have

a significantly lower Jitter (Less Female Mdn: 0.035 (IQR: 0.02 - 0.05); Equal Female Mdn: 0.032 (IQR: 0.02 - 0.05); $p<0.001$, $\eta^2 = 0.01$), Shimmer (Less Female Mdn: 1.50 (IQR: 1.29 - 1.80)); Equal Female Mdn: 1.44 (IQR: 1.20 - 1.75); $p<0.001$, $\eta^2 = 0.01$) and Hammarberg Index (Less Female Mdn: 15.14 (IQR: 7.28 - 20.91); Equal Female Mdn: 11.50 (IQR: 4.23 - 18.20); $p<0.001$, $\eta^2 = 0.03$).

Due to the effect of gender on acoustics, we decided to perform separate LME analysis for male and female participants to remove the effect of gender on the acoustic features. Results from the LME models indicate a similar trend, with the exception of the fundamental frequency with the acoustic features being significantly affected by the ratio of females present in the team as well as how far into the experiment they are. This is expanded further in Table 2.2. We observe a significant positive correlation in the loudness and Spectral Flux for female participants in teams with an equal amount of female participants. The fundamental frequency - both mean and standard deviation - is also significantly affected by how far into the experiment we go.

2.3 Discussion

The results of the analysis reveal significant differences in self-reported experiences, perceptions, and vocal characteristics based on both gender and team composition. These findings provide valuable insights into how gender dynamics may shape individual and team-level outcomes, influencing not only emotional experiences, conflict resolution, and team cohesion but also linguistic and acoustic patterns in group interactions.

Male participants had significantly higher reported PANAS positive scores than the female participants, indicating a greater tendency to experience more positive emotions. This was accompanied by significantly higher perceived stress, indicating that while male participants may report experiencing more positive emotions, they also report greater stress levels. Male participants also reported significantly higher task conflict, implying more frequent disagreements about task execution. Interestingly, conflict management was also significantly higher for male participants, suggesting that despite experiencing greater task conflict, they may engage more in conflict resolu-

Table 2.2: Coefficients and p-values of the linear mixed effects (LME) models

Acoustic Measure	**Ratio of Females (0: Less/1: Equal)**	**Day (1-5)**
F0 Mean	$\beta_1 = 0.902, p = 0.699$	$\boldsymbol{\beta_2 = 0.241, p = 0.001}$
F0 Std	$\beta_1 = 0.456, p = 0.512$	$\boldsymbol{\beta_2 = 0.123, p = 0.045}$
Loudness Mean	$\boldsymbol{\beta_1 = 0.267, p = 0.029}$	$\beta_2 = 0.005, p = 0.388$
Loudness Std	$\beta_1 = 0.189, p = 0.315$	$\beta_2 = -0.017, p = 0.089$
Spectral Flux	$\boldsymbol{\beta_1 = 0.195, p = 0.004}$	$\beta_2 = 0.005, p = 0.213$
MFCC1	$\beta_1 = -3.798, p = 0.352$	$\beta_2 = 0.107, p = 0.439$
MFCC2	$\beta_1 = -2.791, p = 0.447$	$\beta_2 = -0.123, p = 0.397$
MFCC3	$\beta_1 = 1.167, p = 0.759$	$\beta_2 = -0.225, p = 0.120$
MFCC4	$\boldsymbol{\beta_1 = -11.033, p = 0.012}$	$\boldsymbol{\beta_2 = -0.458, p = 0.001}$
Alpha Ratio	$\beta_1 = -1.460, p = 0.523$	$\beta_2 = -0.117, p = 0.330$
Hammarberg Index	$\beta_1 = 1.504, p = 0.624$	$\beta_2 = -0.004, p = 0.981$
Jitter	$\beta_1 = -0.001, p = 0.944$	$\beta_2 = 0.000, p = 0.745$
Shimmer	$\beta_1 = -0.111, p = 0.355$	$\beta_2 = -0.003, p = 0.671$

(a) Female Participants

Acoustic Measure	**Ratio of Females (0: Less/1: Equal)**	**Day (1-5)**
F0 Mean	$\beta_1 = 2.123, p = 0.081$	$\boldsymbol{\beta_2 = -0.178, p = 0.000}$
F0 Std	$\beta_1 = 0.007, p = 0.852$	$\boldsymbol{\beta_2 = 0.005, p = 0.000}$
Loudness Mean	$\beta_1 = 0.071, p = 0.569$	$\boldsymbol{\beta_2 = -0.014, p = 0.000}$
Loudness Std	$\beta_1 = -0.023, p = 0.819$	$\boldsymbol{\beta_2 = 0.013, p = 0.000}$
Spectral Flux	$\beta_1 = -0.040, p = 0.712$	$\boldsymbol{\beta_2 = -0.008, p = 0.000}$
MFCC1	$\boldsymbol{\beta_1 = -7.525, p = 0.008}$	$\beta_2 = -0.001, p = 0.988$
MFCC2	$\beta_1 = -4.743, p = 0.168$	$\boldsymbol{\beta_2 = 0.440, p = 0.000}$
MFCC3	$\beta_1 = 2.652, p = 0.384$	$\boldsymbol{\beta_2 = -0.502, p = 0.000}$
MFCC4	$\beta_1 = -6.139, p = 0.055$	$\boldsymbol{\beta_2 = -0.772, p = 0.000}$
Alpha Ratio	$\beta_1 = 2.717, p = 0.350$	$\beta_2 = -0.099, p = 0.142$
Hammarberg Index	$\beta_1 = -1.821, p = 0.607$	$\beta_2 = 0.017, p = 0.832$
Jitter	$\beta_1 = 0.000, p = 0.952$	$\beta_2 = -0.000, p = 0.425$
Shimmer	$\beta_1 = -0.030, p = 0.764$	$\beta_2 = 0.007, p = 0.138$

(b) Male Participants

tion strategies. Furthermore, female participants reported significantly lower perceived performance and team satisfaction, indicating a lower sense of efficacy and fulfillment in their teamwork experience. Team viability, which reflects how sustainable and promising the team was for the future, was significantly higher among female participants, suggesting that females may perceive their teams as more cohesive and capable of long-term collaboration, despite the lower perceived performance.

Examining the effects of gender composition on team dynamics, we observe that teams with

an equal ratio of male and female members reported significantly lower relationship conflict and task conflict, as well as much higher team cohesion and team viability, as compared to teams with less female members. These findings suggest that gender-balanced teams may promote a more harmonious and collaborative working environment, with reduced interpersonal friction and improved long-term sustainability, which is in line with research by Woolley et. al.[34]. On the other hand, teams with fewer female members reported significantly higher perceived performance and team satisfaction. This could indicate that teams with a male-dominated composition perceive their productivity and collaboration to be stronger, although the underlying reasons for this perception require further investigation.

Teams with fewer female members exhibited significantly higher usage of words associated with analytical thinking, clout, and authenticity compared to gender-balanced teams. The higher prevalence of clout-related words implies that these teams may have stronger expressions of authority. Meanwhile, the higher authenticity scores indicate that male-heavy teams may also express themselves in a more direct or self-revealing manner. On the other hand the significant effect of the female-to-male ratio on tone suggests that the presence of more female members may contribute to a noticeable shift in the overall mood of discussions. Similarly, the significant effect on neutral sentiment implies that gender-balanced teams may exhibit a more moderated or neutral conversational style, possibly leading to more balanced discussions.

The acoustic results reveal notable differences between teams based on the ratio of female members, suggesting that gender composition influences vocal characteristics in group interactions. Specifically, teams with an equal ratio of female members exhibit higher mean fundamental frequency (F0), greater loudness, and distinct spectral properties, including significant variations in multiple MFCCs. The heightened Alpha Ratio and lower Jitter and Shimmer in these teams further support the idea that their speech is clearer and more projected, possibly due to more active engagement and balanced participation among members. Due to the observed effects of gender on acoustics, we performed separate LME analyses for male and female participants to account for the impact of gender on the acoustic features. The results from the LME models indicate a

similar trend across both genders, with the exception of the fundamental frequency. The increased loudness of female speakers in teams that are equally distributed might point to an increase in confidence and participation. Additionally, the LME models indicate that acoustic features evolve over time, with significant effects of experiment duration on F0 variation, loudness, and specific MFCCs. This suggests that vocal patterns adapt throughout the group interaction process.

2.3.1 Limitations

While the findings of this study offer valuable insights into the influence of gender composition on team dynamics, there still exist several limitations that should be acknowledged. First, the reliance on self-reported data may introduce bias, as participants' perceptions of emotions, stress, and conflict could be influenced by personal biases as well as social desirability. on the other hand the acoustic analysis, while informative, may not account for all factors influencing vocal characteristics, such as individual and sociocultural differences in speaking style or environmental influences. Further research also needs to be conducted on the distinct roles of the participants in the study to obtain a more detailed understanding of the role of gender in these interaction.

2.4 Conclusion

This study highlights the significant role that gender composition plays in shaping both individual and team-level outcomes, with important implications for group dynamics. Our findings suggest that gender-balanced teams tend to experience lower levels of conflict, higher cohesion, and stronger team viability, which may promote more harmonious and sustainable collaborations. In contrast, male-dominated teams reported higher perceived performance and satisfaction, though this perception may reflect underlying differences in group interaction and communication dynamics that warrant further investigation. Overall, these findings contribute to our understanding of how gender dynamics influence group processes, from emotional engagement to vocal characteristics. Future research could further explore the mechanisms underlying these patterns, particularly in relation to leadership, conflict resolution, and task performance. Additionally, examining how

these dynamics evolve over time and across different team contexts could offer deeper insights into optimizing team composition for effective collaboration and sustained success.

Bibliography

[1] Umut Avci and Oya Aran. Predicting the performance in decision-making tasks: From individual cues to group interaction. IEEE Transactions on Multimedia, 18, 2016.

[2] Maura J Borrego, Miguel A Padilla, Guili Zhang, Matthew W Ohland, and Timothy J Anderson. Graduation rates, grade-point average, and changes of major of female and minority students entering engineering. In Proceedings Frontiers in Education 35th Annual Conference, pages T3D–1. IEEE, 2005.

[3] Ryan L Boyd, Ashwini Ashokkumar, Sarah Seraj, and James W Pennebaker. The development and psychometric properties of liwc-22. Austin, TX: University of Texas at Austin, 10, 2022.

[4] Bettina Lankard Brown. Women and minorities in high-tech careers. ERIC Clearinghouse on Adult, Career, and Vocational Education, Center on . . . , 2001.

[5] Carlos Busso, Sungbok Lee, and Shrikanth Narayanan. Analysis of emotionally salient aspects of fundamental frequency for emotion detection. IEEE TASLP, 17, 2009.

[6] Jane Carthey, Marc R de Leval, and James T Reason. The human factor in cardiac surgery: errors and near misses in a high technology medical domain. The Annals of thoracic surgery, 72, 2001.

[7] Michael de José Belzunce. Micro-videos and micro-behaviors as an innovative methodology for training in soft skills. In CUICIID 2018. Fórum XXI, 2018.

[8] Roger D Dias, Marco A Zenati, Heather M Conboy, Lori A Clarke, Leon J Osterweil, George S Avrunin, and Steven J Yule. Dissecting cardiac surgery: A video-based recall protocol to elucidate team cognitive processes in the operating room. Annals of surgery, 274, 2021.

[9] Molla S Donaldson, Janet M Corrigan, and Linda T Kohn. To err is human: building a safer health system. 2000.

[10] Antoine Duclos, Michelle L Frits, Christine Iannaccone, Stuart R Lipsitz, Zara Cooper, Joel S Weissman, and David W Bates. Safety of inpatient care in surgical settings: cohort study. bmj, 387, 2024.

[11] Andrew W ElBardissi, Douglas A Wiegmann, Sarah Henrickson, Rishi Wadhera, and Thoralf M Sundt III. Identifying methods to improve heart surgery: an operative approach and strategy for implementation on an organizational level. European Journal of Cardio-Thoracic Surgery, 34, 2008.

[12] Lucca Eloy, Angela EB Stewart, Mary Jean Amon, Caroline Reinhardt, Amanda Michaels, Chen Sun, Valerie Shute, Nicholas D Duran, and Sidney D'Mello. Modeling team-level multimodal dynamics during multiparty collaboration. In 2019 ICMI, 2019.

[13] Susan Geller Ettenheim, Roberta Furger, Lisa Siegman, and Susan McLester. Tips for getting girls involved. Technology & Learning, 20(8):34–36, 2000.

[14] Florian Eyben, Klaus R Scherer, Björn W Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y Devillers, Julien Epps, Petri Laukka, Shrikanth S Narayanan, et al. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. IEEE TAffC, 7, 2015.

[15] Richard M Felder, Gary N Felder, and E Jacquelin Dietz. The effects of personality type on engineering student performance and attitudes. Journal of engineering education, 91(1):3–17, 2002.

[16] Stine Gundrosen, Ellen Andenæs, Petter Aadahl, and Gøril Thomassen. Team talk and team activity in simulated medical emergencies: a discourse analytical approach. Scandinavian journal of trauma, resuscitation and emergency medicine, 24, 2016.

[17] Veena Guru, Jack V Tu, Edward Etchells, Geoffrey M Anderson, C David Naylor, Richard J Novick, Christopher M Feindel, Fraser D Rubens, Kevin Teoh, Avdesh Mathur, et al. Relationship between preventability of death after coronary artery bypass graft surgery and all-cause risk-adjusted mortality rates. Circulation, 117, 2008.

[18] John HL Hansen and Sanjay Patil. Speech under stress: Analysis, modeling and recognition. Speaker classification I: Fundamentals, features, and methods, 2007.

[19] James Hillenbrand, Laura A Getty, Michael J Clark, and Kimberlee Wheeler. Acoustic characteristics of american english vowels. The Journal of the Acoustical society of America, 97(5):3099–3111, 1995.

[20] Swathi Jagannath, Aleksandra Sarcevic, and Ivan Marsic. An analysis of speech as a modality for activity recognition during complex medical teamwork. In 12th EAI International conference on pervasive computing technologies for healthcare, 2018.

[21] Gideon Keren and Charles Lewis. A handbook for data analysis in the behavioral sciences: Methodological issues. L. Erlbaum Associates Hillsdale, NJ, 1993.

[22] Martin A Makary and Michael Daniel. Medical error—the third leading cause of death in the us. Bmj, 353, 2016.

[23] Tamara J Moore, Micah S Stohlmann, Hui Hui Wang, Kristina M Tank, Aran W Glancy, and Gillian H Roehrig. Implementation and integration of engineering in k-12 stem education. In Engineering in pre-college settings: Synthesizing research, policy, and practices, pages 35–60. Purdue University Press, 2014.

[24] Gabriel Murray and Catharine Oertel. Predicting group performance in task-based interaction. In Proceedings of the 20th ACM ICMI, 2018.

[25] NC NCSES. Diversity and stem: Women, minorities, and persons with disabilities 2023, 2023.

[26] José C Pinheiro and Douglas M Bates. Linear mixed-effects models: basic concepts and examples. Mixed-effects models in S and S-Plus, 2000.

[27] Joseph M Reilly and Bertrand Schneider. Predicting the quality of collaborative problem solving through linguistic analysis of discourse. International Educational Data Mining Society, 2019.

[28] Jonathan Svensson and Jan Andersson. Speech acts, communication problems, and fighter pilot team performance. Ergonomics, 49, 2006.

[29] Yla R Tausczik and James W Pennebaker. The psychological meaning of words: Liwc and computerized text analysis methods. Journal of language and social psychology, 29, 2010.

[30] Karen L Tonso. Teams that work: Campus culture, engineer identity, and social interactions. Journal of engineering education, 95(1):25–37, 2006.

[31] Joyce A Wahr, Richard L Prager, JH Abernathy Iii, Elizabeth A Martinez, Eduardo Salas, Patricia C Seifert, Robert C Groom, Bruce D Spiess, Bruce E Searles, Thoralf M Sundt III, et al. Patient safety in the cardiac operating room: human factors and teamwork: a scientific statement from the american heart association. Circulation, 128, 2013.

[32] Joanna Wolfe and Kara Poe Alexander. The computer expert in mixed-gendered collaborative writing groups. Journal of Business and Technical Communication, 19(2):135–170, 2005.

[33] Joanna Wolfe and Elizabeth Powell. Biases in interpersonal communication: How engineering students perceive gender typical speech acts in teamwork. Journal of Engineering Education, 98(1):5–16, 2009.

[34] Anita Williams Woolley, Christopher F Chabris, Alex Pentland, Nada Hashmi, and Thomas W Malone. Evidence for a collective intelligence factor in the performance of human groups. science, 330(6004):686–688, 2010.