Predicting User Confidence in Video Recordings with Spatio-Temporal Multimodal Analytics

Andrew Emerson
aemerson@ets.org
AI Research Labs, Educational Testing Service
Princeton, New Jersey, USA

Patrick Houghton
phoughton@ets.org
AI Research Labs, Educational Testing Service
Princeton, New Jersey, USA

Ke Chen
ck19970920@gmail.com
AI Research Labs, Educational Testing Service
Princeton, New Jersey, USA

Vinay Basheerabad
vbasheerabad@ets.org
AI Research Labs, Educational Testing Service
Princeton, New Jersey, USA

Rutuja Ubale
rubale@ets.org
AI Research Labs, Educational Testing Service
Princeton, New Jersey, USA

Chee Wee Leong
cleong@ets.org
AI Research Labs, Educational Testing Service
Princeton, New Jersey, USA

ABSTRACT

A critical component of effective communication is the ability to project confidence. In video presentations (e.g., video interviews), there are many factors that influence perceived confidence by a listener. Advances in computer vision, speech processing, and natural language processing have enabled the automatic extraction of salient features that can be used to model a presenter's perceived confidence. Moreover, these multimodal features can be used to automatically provide feedback to a user with ways they can improve their projected confidence. This paper introduces a multimodal approach to modeling user confidence in video presentations by leveraging features from visual cues (i.e., eye gaze) and speech patterns. We investigate the degree to which the extracted multimodal features were predictive of user confidence with a dataset of 48 2-minute videos, where the participants used a webcam and microphone to record themselves responding to a prompt. Comparative experimental results indicate that our modeling approach of using both visual and speech features are able to score 83% and 78% improvements over the random and majority label baselines, respectively. We discuss implications of using the multimodal features for modeling confidence as well as the potential for automated feedback to users who want to improve their confidence in video presentations.

CCS CONCEPTS

• **Human-centered computing**; • **Computing methodologies** → **Machine learning**;

KEYWORDS

multimodal analytics, confidence measurement, audio-visual

ICMI '22 Companion, November 7–11, 2022, Bengaluru, India

ACM ISBN 978-1-4503-9389-8/22/11…$15.00
https://doi.org/10.1145/3536220.3558007

ACM Reference Format:
Andrew Emerson, Patrick Houghton, Ke Chen, Vinay Basheerabad, Rutuja Ubale, and Chee Wee Leong. 2022. Predicting User Confidence in Video Recordings with Spatio-Temporal Multimodal Analytics. In *INTERNATIONAL CONFERENCE ON MULTIMODAL INTERACTION (ICMI '22 Companion), November 7–11, 2022, Bengaluru, India.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3536220.3558007

1 INTRODUCTION

Confidence plays a central role in effective communication [4, 20]. The ability to confidently express a point, recount a story, describe one's skills or experience, or relay an important message has a significant impact on a listener's perception and understanding. Projecting confidence effectively is a skill that requires practice and feedback, but actionable feedback to improve one's confidence requires a comprehensive understanding of the speaker's communication. One way of practicing projecting confidence in speech is through video recordings of oneself responding to a prompt. Due to the complexity of confidence as a construct, many factors such as visual cues (e.g., facial expressions, eye movements) and speech patterns (e.g., speaking rate) have synergistic effects that influence a listener's perceived confidence of the speaker. Recent advances in computer vision, speech processing, and natural language processing have afforded the automatic, multimodal analysis and detection of confidence of humans in video recordings.

Inducing computational models of user confidence poses several challenges. Due to cultural, gender, and personality differences, people may express confidence differently within a given scenario (e.g., responding to a prompt in a video). As a result, it is critical to derive fair, converged labels of confidence from a diverse set of annotators. This challenge demands an extensive, careful solution to labeling that often yields a smaller set of high-quality labeled data. With a smaller dataset, automated models of confidence that leverage salient sets of multimodal data must be able to accurately identify patterns of confidence or a lack thereof despite the sample size. To address these challenges, it is critical to devise a principled rubric to label confidence that is applied by a set of annotators, who come to a consensus agreement. The models trained from the subsequent labeled data must also be able to detect unique patterns related to the multimodal cues stemming from visual and verbal (vocal) features.

In this paper, we introduce a multimodal analytics framework for predicting user confidence in 2-minute video presentations, where each user responds to a common prompt. The analysis used data collected from a set of 1891 videos, where a small subset of 48 videos were carefully annotated following the creation of a rubric. We generated a set of multimodal features derived from both the visual and verbal modalities that are intended to identify human-interpretable patterns that influence how a listener perceives a speaker's confidence level. We conducted a series of ablation experiments comparing the performance of predictive models that use each possible set of modalities to both unimodal baselines and baselines that do not use the engineered features. This is among the first work to use multimodal data to capture patterns of user confidence, and we discuss the implications of how these findings can be used to provide constructive feedback to users about their projected confidence.

2 LITERATURE REVIEW

In this study, we seek to identify multimodal cues for speaker confidence to which a listener might attune, and we model those cues in order to provide automated feedback to speakers about how they may be perceived. Our immediate aim is to utilize only visual and audio signals, without considering the content of speech. While the construct of *perceived confidence* may seem subjective and difficult to pin down, there is evidence that multiple visual and auditory expressions of confidence are implicitly recognized by observers, and that conclusions based on those expressions are made quickly and are subconsciously used to make decisions.

When observing or listening to a speaker, many different streams of information are combined in order to arrive at conclusions about that speaker. The individual streams, though, also carry enough information to inform decisions [2]. Neurological studies show that there are multiple nonverbal aspects of communication active in both production and perception of speech that allow for conveyance of information beyond the content of language [8], and that differential conclusions based on that information happen quickly [14, 15]. This nonverbal information has been shown to impact high stakes decision-making about a speaker, including hireability decisions [3, 6]. Specific to our concerns here, the perception of a speaker's confidence can be influenced by particular visual and nonverbal audio cues. A recent investigation showed that a difference just in a speaker's tone (confident, doubtful, or neutral voice) could influence a speaking partner's level of trust in a game scenario [4]. In another study, where audio cues were controlled, an investigation showed that observers could accurately perceive confidence based on visual cues alone [20]. This work from across multiple disciplines shows that meaningful conclusions and decision-making processes related to a speaker's perceived confidence are influenced by each modality individually, as well as in toto. Given these conditions, we aim to develop a baseline system for predicting perceived confidence that utilizes visual and auditory cues, while leaving the contribution of the content of speech for later work.

Few studies have utilized multimodal data to model confidence, but there are many studies that model other related constructs, such as personality. One investigation proposed a multimodal framework to predict Big Five personality traits [23], where results indicated that an approach that captures temporal relationships of audio-visual cues outperforms baselines that consider time segments independently. Moreover, this study found that when compared to the audio- or video-only unimodal settings, audio-visual annotations and visual cues yield the best prediction results collectively. Personality prediction has been leveraged in the area of hiring, where one study developed a multimodal system to automatically screen job applicant's personality traits from short video presentations to inform whether to invite the candidate for an in-person interview [9]. Another study found that for predicting Big Five personality traits in a hiring setting, fusing all possible audio-visual modalities obtained the highest performance [11]. Recent work has found that deep learning approaches such as long short-term memory (LSTM) networks and convolutional neural networks (CNNs) are effective for predicting personality traits with multimodal data [13]. We build upon these related findings for predicting confidence by incorporating audio-visual features that are extracted from user videos.

3 VIDEO DATASET

The video dataset used for this study was collected in prior work [5], and consists of responses to past-focused structured interview prompts (e.g., see Table 1), collected via a browser-based video recording platform. Participants recorded themselves responding to up to eight prompts using their computer's webcam and microphone. Participants were recruited through Amazon Mechanical Turk and were limited to respondents within the United States. The pool of participants represents a diverse set of gender, race, and age. The set consists of 1891 videos, each presenting a single response to one of eight prompts from one of 260 total participants (not all participants responded to every prompt).

Question: Please tell us about a work situation in which you were not the formal leader but tried to assume a leadership role. Please provide: • details about the background of the situation, • the behaviors you carried out in response to that situation, • and what the outcome was.

Table 1: Example prompt from [5]

3.1 Confidence Annotation

The confidence annotations are one of the main contributions of this paper, and none of the labels from previous work on the dataset [16–18] have been utilized for the current study. For an initial exploratory annotation effort, a random sample of 50 videos was selected, with two being removed from the set due to audio flaws. This sample of 48 videos was culled from the full set of 1891 videos, with no restriction on participant or prompt (i.e., many of the prompts were represented). Among the sampled video participants, there are 20 males and 28 females while the racial diversity covers 32 White, 11 Black/African-American, 4 Asian/Asian-American and 1 Hispanic/Latino participants. A larger, crowd-sourced annotation is underway using all responses to a single prompt (the prompt in Table 1), for a total of 223 videos. This approach was chosen to ensure diversity of the participants, as each respondent could be represented only once, as well as to ensure uniformity of topic,

Label	Description
1	**Lower confidence** speakers may exhibit the following: hesitation between and within sentences; self-correction and/or false starts; reserved voice quality; closed or "protective" posture; fidgeting or extraneous body movement; frequent changes in focal point or consistent downward focus of eye gaze
2	**Average confidence** speakers may exhibit the following: limited hesitation; fidgeting is limited; voice quality is comfortable and conversational with few false-starts; posture is open and toward screen/camera; eye gaze mostly focused on screen/camera
3	**Higher confidence** speakers may exhibit the following: body posture and facial expression appear comfortable; intentional gesturing; does not focus on errors (correct and/or move on); voice quality is comfortable or projected (as appropriate for a presentation); attentive, upright body posture; consistent eye gaze aimed at screen or camera

Table 2: Confidence rubric

avoiding any unexpected variation in responses due to prompt subject.

During the first pass annotation of the exploratory 48-video set, videos were first labeled by one annotator on a five-point scale. However, when a subset was labeled by a second annotator, the inter-annotator agreement was not acceptable, likely due to the somewhat subjective nature of the task. The rubric was then collapsed to a three-point scale and all further annotation was done with this scale. The label descriptions used for the three-point scale are shown in Table 2. Using the three-point scale, two raters independently annotated each video, with quadratic weighted kappa agreement of 0.65. For the videos where they disagreed, the raters came to an agreement after discussion. For instances where an agreement could not be reached (three total cases), a third rater labeled the video for the final rating. The final distribution of labels for this set was: Low (label 1) = 13; Average (label 2) = 16; High (label 3) = 19.

The larger, crowd-sourced annotation effort on 223 videos is underway, but we have not yet begun to use those data for our models. For these annotations, 10 annotators scored each video, and the consensus score was reported. Consensus was determined by weighted voting, where annotators who perform better on calibration video samples inserted into the data would receive a greater weight to their votes. While initial agreement on the annotations is good (over 80% agreement), the distribution of scores is overwhelmingly weighted toward the central value (83% of consensus responses were 'average confidence', or 2 on the scale in Table 2). As a result, we are currently finetuning the rubric to help annotators distinguish between categories. Further investigation is needed before using this larger set of data.

4 MULTIMODAL FEATURE EXTRACTION

4.1 Speech Modality

To extract features from audio captured by the speaker's microphone (i.e., the speech modality), we deployed an in-house speech engine built using Apache Storm[1]. The engine processes audio, transcribes the audio, and generates statistics that summarize the speech. The statistics computed by the engine depend on the service requested. Specifically, the engine computes several basic speech fluency features with most of the features coming from the in-house scoring engine for spoken responses. In total, we extracted six features from the speakers' audio, described in Table 3.

[1]https://storm.apache.org/

4.2 Visual Modality

To extract features captured by the user's webcam (i.e., the visual modality), we processed each frame from the videos using MediaPipe Face Mesh [1, 10]. Face Mesh is a toolkit that approximates 468 3D facial landmarks and an additional 10 landmarks outlining the pupils and irises. We calculated two set of features derived from the output of MediaPipe Face Mesh to characterize user behavior from the visual perspective.

First, we engineered a set of human-interpretable statistical features to enable actionable feedback to the user. To extract meaningful features from the facial landmarks, we selected landmarks related to eye gaze and head posture. Specifically, we selected the landmarks for each pupil, the left-most and right-most positions of the face, and the top and bottom of the face. Using the pupil locations, we approximated the angle that the user was looking, both in the vertical axis (i.e., y-axis) and horizontal axis (i.e., x-axis). We also computed the roll, yaw, and pitch angles of the users' face to achieve a comprehensive measurement of the user's head pose. For head pose, we converted the angles into a single value denoting the displacement from the center of the screen at an upright angle. Using these calculations for each frame in the video, we computed the mean, variance, kurtosis, and skew of the Euclidean distances of the current frame from both the origin (center of the screen) and the previous frame's position. The four statistical values for the origin and previous frame were computed both for eye gaze and head pose, for a total of 16 statistical features.

Second, we leveraged the sequential nature of the landmark data by recording the pupil location in each frame. This resulted in a sequence of three values: the X, Y, and Z coordinate locations of the pupil at each time point or frame. To avoid redundancy, we only recorded the locations of one pupil with the assumption that both eye movements are symmetrical. Given that eye movements were helpful in previous research to model and provide inference of mental states in humans [12], we hypothesize that the number of eye anomalies exhibited (e.g., sudden, atypical eye movements) are potentially useful features for prediction of confidence. In our approach, we first used a CNN autoencoder to learn a latent representation, $\boldsymbol{h}$, of the input sequence data, $\boldsymbol{x}$, such that it can be used to reconstruct the input. Formally, given an encoder function $\boldsymbol{f(x)}$ and a decoder function $\boldsymbol{g(x)}$, we performed multiple iterations of training over the full set of sequence data (1843 videos) in order

Feature	Description
Response Length	number of content words (not silence, noise, and hesitation markers) in the ASR hypothesis
Silences per Token	number of silences / number of tokens (including silences, hesitation markers, repetitions, and content words), where silences shorter than 0.145 seconds are excluded
Pauses Ratio	total pause duration / total speaking time, where pauses include silences (>= 0.145 seconds) and hesitation markers
Number of Hesitation Markers	number of "um", "uh", and similar words
Filler Ratio	total number of disfluencies / total number of content words
Speaking Rate	number of content words in the response / total utterance duration

Table 3: Verbal features set and their corresponding descriptions

to minimize the mean squared error loss, defined as $L(\mathbf{x}, g(f(\mathbf{x}))$. Specifically, our learned latent representation has dimension $|\boldsymbol{h}|<|\boldsymbol{x}|$ that resulted in an undercomplete autoencoder, while $\boldsymbol{f}$ and $\boldsymbol{g}$ are nonlinear functions with regularization achieved by dropout, hence affording a more powerful generalization of other dimension reduction techniques, such as PCA. The capacity of the network is defined by the number of filters in the cascaded convolution and deconvolution layers to be 32, 16, 16, and 32, respectively. For each layer, we used a kernel with size 7 and a stride with size 2 with padding. The ReLU activation function is used in all layers.

In order to avoid data leakage, note that the full sequence data used to train the autoencoder is not used for any of our model building and evaluation in our experiments. Given that our 2-minute videos are converted to 30 FPS for experimentation, there are a total number of 3600 frames extracted per video. Prior to running each epoch, we applied a max pooling over the frames by using a block size of (30,1) to reduce the time-series data to 120 time intervals, where each interval is equivalent to one second over which the X, Y or Z coordinate value is extracted. After training the autoencoder for 100 epochs each for the three dimensional values, the MSE loss, L, converged at < 0.05.

Our anomaly detection is based on the premise: If a spatio-temporal gaze pattern is *atypical* and not observed in our training data, the autoencoder will reconstruct such patterns with high error rates. Specifically, an anomaly is counted as such when the mean absolute error between the reconstructed signal and the original input exceeds the threshold, which is computed as the mean of the signal + one standard deviation obtained from the training data. For each video, we extracted the number of anomalies in the X, Y, and Z dimensions to be used as inputs for predictive models.

5 MODELING USER CONFIDENCE

Following our multimodal feature extraction, we group all features by their originating modality, and aggregate features into one of the *verbal* or *visual* feature sets. Note that, in a true multimodal approach, the textual modality is typically included, but we have purposely omitted text to afford a confidence prediction model is content-agnostic.

We conducted two experiments to evaluate the efficacy of the modality-based feature sets, as well as the contribution by each individual feature to overall confidence modeling. In the first experiment, modality-specific features are used as inputs to four popular classifiers known for their effectiveness across a wide variety of ML tasks [7, 19, 21, 22], namely Logistic Regression, SVC, Decision Tree and Random Forest. For each model, we performed grid search across several of the hyperparameter choices, and for others, we held constant (e.g., RBF kernel for the SVC, L2 loss for the Logistic Regressor). Next, the modality-specific features are fused together at the feature-level for a bimodal approach where we used all nine features for modeling confidence. Evaluation is reported using three metrics: F1 (macro-averaged across all classes), Accuracy, and Quadratic Weighted Kappa, where each metric targets a specific performance aspect of the models. The results are illustrated in Table 4. The * notation indicates the model statistically significantly outperforms a baseline (i.e., either Random or Majority Label), and the ** notation indicates the model outperforms both baselines for the given metric. The comparisons between models were conducted using the one-sided Wilcoxon signed-rank test, which compares the performance of the models per cross-validation fold.

Due to the use of the dataset of 48 videos for evaluation, we refrain from performing correlations between the verbal or visual features and the labels prior to the experiments to avoid data leakage, since all 48 videos are used, at some point, as a datapoint in a 10-fold cross-validation. Instead, in the second experiment, we conducted ablation where each feature (or feature set) is removed, one at a time, from the best-performing model configuration (i.e., Random Forest learner using a bimodal approach in Table 4), to observe if such a removal would result in a drop in model performance.

Previously, we described the extraction of frame-based head pose and eye gaze features and their statistical characteristics (mean, variance, kurtosis, and skew) of the Euclidean distances of the current frame from both the origin (center of the screen) and the previous frame's position. These time-aggregated visual features were included in our preliminary experimentation, but generated inconclusive results when compared to modeling using eye gaze anomalies. Due to space constraints, we have omitted their reporting and analysis from Tables 4 and 5 but plan to revisit them for future work. Importantly, our intended use case for confidence modeling involves providing actionable and interpretable feedback to users. Hence, the time-aggregated visual features, which could potentially improve the model performance, might eventually be deprioritized due to a lack of clarity in how they work.

The results in Table 4 reveal several interesting patterns. While combining both verbal and visual feature sets yielded the best performing system across the three metrics (i.e., F1 = 0.705 using SVC,

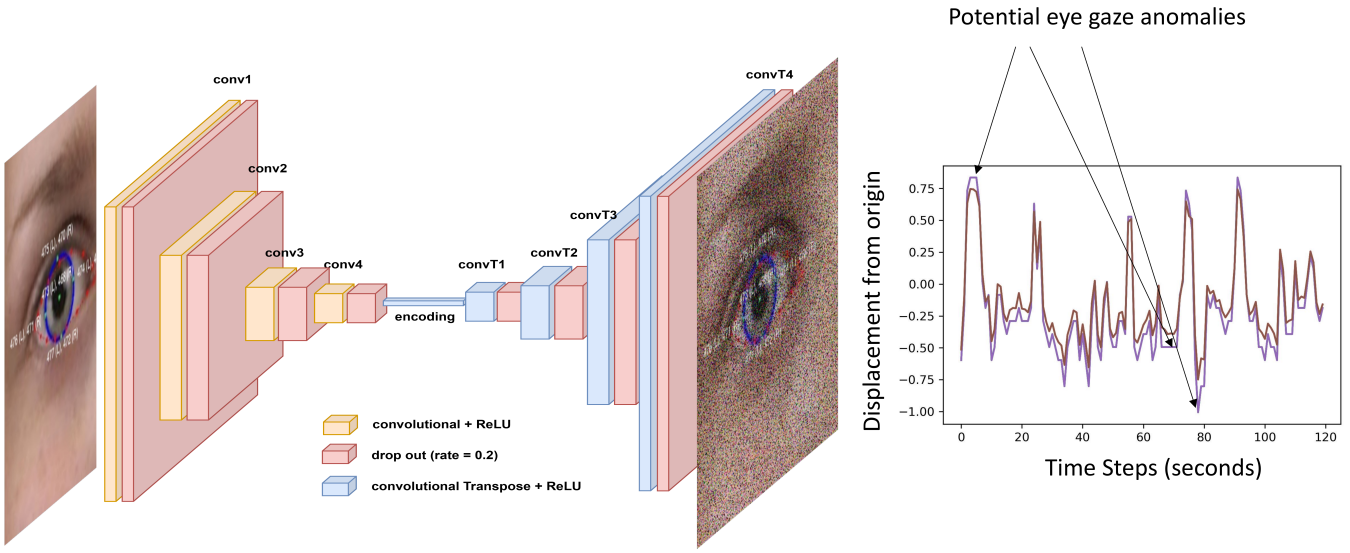

Figure 1: Diagram (left) showing a *generalized* autoencoder with symmetrical convolutional (encoding) and convolutional-transpose (decoding) layers. The number of filters decreases when moving toward the encoding (or code), while it increases moving away from the encoding, while kernel size and stride are typically kept constant. The graph (right) shows a plot of the original input sequence data (red) and the reconstructed sequence data (purple). Discrepancies observed are potentially eye gaze anomalies when the MAE is more than a threshold (mean + 1 std) computed from the training data. Anomalies are computed independently for X, Y and Z directions of the eye gaze.

Acc. = 0.715 using Random Forest, QW_k = 0.762 using Random Forest), using exclusively the verbal feature set allows us to construct a model that significantly outperforms the random and majority baselines. This is perhaps not surprising given past findings in the literature about the usefulness of speech-related features in modeling social, behavioral constructs [22]. Given the novelty of the eye gaze anomaly features, we are particularly interested in a confidence model that employs exclusively such features. By themselves, they are not demonstrably effective, with the best configuration (F1 = 0.334, Accu. = 0.390, QW_k = 0.128, all using Logistic Regression). However, the eye gaze anomalies-based model is perhaps mining a different pattern from the data that is not random nor insisting on a fixed label like the majority class. This is evident upon looking at the QW_k metric values where different learners (in the visual modality) either achieve a higher value (e.g., QW_k = 0.128, Logistic Regression) or a negative value (e.g., QW_k = -0.016, SVC) when compared to those of random and majority baselines. Note that, QW_k typically varies from 0 (random agreement between raters) to 1 (complete agreement between raters). In the event that there is less agreement between the raters than expected by chance, the metric may go below 0. However, given more context in the presence of speech features, anomaly features contribute significantly to overall model performance compared to baselines (see Table 4). For the best performing bimodal model (i.e., Random Forest), we note an interesting trend for the performance on specific class labels. Specifically, the F1 score is much higher on classes 1 and 3 (0.817 and 0.780) compared to a lower performance on class 2 (0.433). This indicates the model is much better at identifying extreme ends of confidence or lack of confidence, but often misidentifies moderate levels.

In the ablation study findings reported in Table 5, we observe that all features (or feature sets), when removed, resulted in a drop in performance of the Random Forest-based, best performing bimodal system, across all metrics, with the exception of the filler ratio and speaking rate for F1. This perhaps suggests that, while those two verbal features are useful for modeling confidence measured by accuracy and QW_k, their absence can be mitigated by other verbal features in a way that more than compensates for their contribution when present, when taking F1 model performance into account. As expected, the largest decrease across performance metrics occurs when all verbal features are dropped. Dropping all visual features also resulted in a decrease across all metrics ranging from 0.08 (QW_k) to 0.12 (Acc.). Overall, all engineered features are helpful when evaluating their contribution using at least one model performance metric.

6 CONCLUSION

Confidence measurement has important applications for effective communication. The ability to automate such a measurement, coupled with informative and actionable feedback, is an important research area for enabling self-coaching. From creating a valid rubric to evaluating predictive models, automating confidence prediction presents challenges as much as it excites. Firstly, we see that human evaluation of confidence is highly subjective, requiring lengthy discussions and arbitration to finalize annotations. As such, while our bimodal model based on verbal and visual cues outperforms the random and majority baselines significantly, it is based on a relatively small evaluation dataset of videos. Crowd sourcing of confidence labels based on a larger set of annotators is currently underway in order to generate a more representative,

Modality	Features	Classifier	F1	Acc.	QW_k
Verbal	response length, #silences per token, pauses ratio, #hesitation markers, filler ratio, speaking rate	Logistic Regression	0.574**	0.640**	0.705**
		SVC	0.444	0.480	0.577
		Decision Tree	0.412	0.470	0.480
		Random Forest	0.522	0.595	0.682
Visual	eye gaze: anomalies (X), anomalies (Y), anomalies (Z)	Logistic Regression	0.334	0.390	0.128
		SVC	0.237	0.330	-0.016
		Decision Tree	0.308	0.350	0.080
		Random Forest	0.250	0.365	0.077
Bimodal	all verbal + visual features	Logistic Regression	0.512	0.605	0.668
		SVC	**0.705****	0.710	0.695
		Decision Tree	0.629	0.705	0.619
		Random Forest	0.677	**0.715****	**0.762****
Random			0.389	0.390	0.091
Majority Label			0.189	0.400	0.000

Table 4: Automated prediction of user's confidence. Models with different feature sets are compared using three metrics in a 10-fold cross-validation setting over the 48-video dataset. * indicates the model outperforms either the random or majority baseline, and ** indicates it outperforms both ($p < 0.05$)

Feature(s)	F1	Change	Acc.	Change	QW_k	Change
- Verbal (all)	0.250	↓	0.365	↓	0.077	↓
- response length	0.651	↓	0.675	↓	0.744	↓
- #silences per token	0.592	↓	0.650	↓	0.678	↓
- pauses ratio	0.544	↓	0.585	↓	0.544	↓
- #hesitation markers	0.660	↓	0.695	↓	0.743	↓
- filler ratio	0.694	↑	0.715	–	0.735	↓
- speaking rate	0.704	↑	0.710	↓	0.746	↓
- Visual (all)	0.522	↓	0.595	↓	0.682	↓
- anomalies (X)	0.675	↓	0.715	–	0.755	↓
- anomalies (Y)	0.635	↓	0.675	↓	0.709	↓
- anomalies (Z)	0.619	↓	0.695	↓	0.728	↓

Table 5: Feature (or modality) ablation experimental results where features (or feature sets) are removed, one at a time, using a Random Forest model in the same 10-fold cross-validation setting. The direction of change in F1, Accuracy and QW_k are with respect to performance of the bimodal system using the Random Forest model in Table 4

consensus-based rating per video. However, there are bias and fairness issues related to gender, race, ethnicity, and social norms that adds complexity to the annotation process, and these must be mitigated. We intend to address those concerns via frameworks such as Fairlearn[2], where we can quantify the level of progress towards partial or full mitigation. The second challenge is perhaps unique to our social mission, where we aspire to present to the users measurable, transparent, and particularly actionable feedback for self-improvement in a longitudinal manner. This runs contrary to end-to-end approaches for generating a summative confidence label, which has its attraction, but requires deep domain expertise to perform feature engineering, ultimately accounting for more time in the product development life cycle. Ultimately, we are just scratching the surface in our exploratory study. For now, to exude confidence, one might include best practices from our initial analyses: looking left-right occasionally has limited negative effects in terms of perceived confidence, but looking up-down frequently tends to appear less confident. As is often recommended when speaking, our findings suggest that avoiding filler words such as "um" and "ah", and pausing momentarily after making a point are advantageous strategies to appear more confident when speaking.

[2]https://fairlearn.org/

REFERENCES

[1] Artsiom Ablavatski, Andrey Vakunov, Ivan Grishchenko, Karthik Raveendran, and Matsvei Zhdanovich. 2020. Real-time Pupil Tracking from Monocular Video for Digital Puppetry. https://doi.org/10.48550/ARXIV.2006.11341
[2] M. Argyle. 1972. Non-verbal communication in human social interaction. *Non-Verbal Communication* (1972).
[3] Julie Brosy, Adrian Bangerter, and Eric Mayor. 2016. Disfluent Responses to Job Interview Questions and What They Entail. *Discourse Processes* 53 (02 2016). https://doi.org/10.1080/0163853X.2016.1150769

[4] Jonathan Caballero and Marc Pell. 2020. Implicit effects of speaker accents and vocally-expressed confidence on decisions to trust. *Decision* 7 (10 2020). https://doi.org/10.1037/dec0000140

[5] Lei Chen, Ru Zhao, Chee Wee Leong, Blair Lehman, Gary Feng, and Mohammed Ehsan Hoque. 2017. Automated video interview judgment on a large-sized corpus collected online. In *2017 Seventh International Conference on Affective Computing and Intelligent Interaction (ACII)*. 504–509. https://doi.org/10.1109/ACII.2017.8273646

[6] Timothy Degroot and Æ Janaki Gooty. 2009. Can Nonverbal Cues be Used to Make Meaningful Personality Attributions in Employment Interviews. *Journal of Business and Psychology* (2009), 179–192.

[7] Andrew Emerson, Nathan Henderson, Jonathan Rowe, Wookhee Min, Seung Lee, James Minogue, and James Lester. 2020. *Early Prediction of Visitor Engagement in Science Museums with Multimodal Learning Analytics*. Association for Computing Machinery, New York, NY, USA, 107–116. https://doi.org/10.1145/3382507.3418890

[8] Sascha Frühholz and Stefan R. Schweinberger. 2021. Nonverbal auditory communication – Evidence for integrated neural systems for voice signal production and perception. *Progress in Neurobiology* 199 (2021), 101948. https://doi.org/10.1016/j.pneurobio.2020.101948

[9] Jelena Gorbova, Iiris Lusi, Andre Litvin, and Gholamreza Anbarjafari. 2017. Automated Screening of Job Candidate Based on Multimodal Video Processing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

[10] Ivan Grishchenko, Artsiom Ablavatski, Yury Kartynnik, Karthik Raveendran, and Matthias Grundmann. 2020. Attention Mesh: High-fidelity Face Mesh Prediction in Real-time. https://doi.org/10.48550/ARXIV.2006.10962

[11] Yağmur Güçlütürk, Umut Güçlü, Xavier Baró, Hugo Jair Escalante, Isabelle Guyon, Sergio Escalera, Marcel A.J. van Gerven, and Rob van Lier. 2018. Multimodal First Impression Analysis with Deep Residual Networks. *IEEE Transactions on Affective Computing* 9, 3 (2018), 316–329. https://doi.org/10.1109/TAFFC.2017.2751469

[12] Stephen Hutt, Caitlin Mills, Nigel Bosch, Kristina Krasich, James Brockmole, and Sidney D'mello. 2017. " Out of the Fr-Eye-ing Pan" Towards Gaze-Based Models of Attention during Learning with Technology in the Classroom. In *Proceedings of the 25th conference on user modeling, adaptation and personalization*. 94–103.

[13] Julio C. S. Jacques Junior, Yağmur Güçlütürk, Marc Pérez, Umut Güçlü, Carlos Andujar, Xavier Baró, Hugo Jair Escalante, Isabelle Guyon, Marcel A. J. van Gerven, Rob van Lier, and Sergio Escalera. 2022. First Impressions: A Survey on Vision-Based Apparent Personality Trait Analysis. *IEEE Transactions on Affective Computing* 13, 1 (2022), 75–95. https://doi.org/10.1109/TAFFC.2019.2930058

[14] Xiaoming Jiang, Kira Gossack-Keenan, and Marc D Pell. 2020. To believe or not to believe? How voice and accent information in speech alter listener impressions of trust. *Quarterly Journal of Experimental Psychology* 73, 1 (2020), 55–79. https://doi.org/10.1177/1747021819865833 arXiv:https://doi.org/10.1177/1747021819865833 PMID: 31293191.

[15] Xiaoming Jiang and Marc D. Pell. 2015. On how the brain decodes vocal cues about speaker confidence. *Cortex* 66 (2015), 9–34. https://doi.org/10.1016/j.cortex.2015.02.002

[16] Chee Wee Leong, Xianyang Chen, Vinay Basheerabad, Chong Min Lee, and Patrick Houghton. 2021. NLP-guided Video Thin-slicing for Automated Scoring of Non-Cognitive, Behavioral Performance Tasks. In *Proceedings of the 2021 International Conference on Multimodal Interaction*. 846–847.

[17] Chee Wee Leong, Katrina Roohr, Vikram Ramanarayanan, Michelle P Martin-Raugh, Harrison Kell, Rutuja Ubale, Yao Qian, Zydrune Mladineo, and Laura McCulla. 2019. Are Humans Biased in Assessment of Video Interviews?. In *Adjunct of the 2019 International Conference on Multimodal Interaction*. 1–5.

[18] Chee Wee Leong, Katrina Roohr, Vikram Ramanarayanan, Michelle P Martin-Raugh, Harrison Kell, Rutuja Ubale, Yao Qian, Zydrune Mladineo, and Laura McCulla. 2019. To trust, or not to trust? A study of human bias in automated video interview assessments. *arXiv preprint arXiv:1911.13248* (2019).

[19] Haley Lepp, Chee Wee Leong, Katrina Roohr, Michelle Martin-Raugh, and Vikram Ramanarayanan. 2020. *Effect of Modality on Human and Machine Scoring of Presentation Videos*. Association for Computing Machinery, New York, NY, USA, 630–634. https://doi.org/10.1145/3382507.3418880

[20] Yondu Mori and Marc D. Pell. 2019. The Look of (Un)confidence: Visual Markers for Inferring Speaker Confidence in Speech. *Frontiers in Communication* 4 (2019). https://doi.org/10.3389/fcomm.2019.00063

[21] Gabriel Murray and Catharine Oertel. 2018. Predicting Group Performance in Task-Based Interaction. In *Proceedings of the 20th ACM International Conference on Multimodal Interaction* (Boulder, CO, USA) *(ICMI '18)*. Association for Computing Machinery, New York, NY, USA, 14–20. https://doi.org/10.1145/3242969.3243027

[22] Chinchu Thomas and Dinesh Jayagopi. 2022. Predicting Presentation Skill of a Speaker Using Automatic Speaker and Audience Measurement. *IEEE Transactions on Learning Technologies* (2022), 1–1. https://doi.org/10.1109/TLT.2022.3171601

[23] Oya Çeliktutan and Hatice Gunes. 2017. Automatic Prediction of Impressions in Time and across Varying Context: Personality, Attractiveness and Likeability. *IEEE Transactions on Affective Computing* 8, 1 (2017), 29–42. https://doi.org/10.1109/TAFFC.2015.2513401