SpeechMirror: A Multimodal Visual Analytics System for Personalized Reflection of Online Public Speaking Effectiveness

Zeyuan Huang, Qiang He, Kevin Maher, Xiaoming Deng, *Member, IEEE*, Yu-Kun Lai, *Member, IEEE*, Cuixia Ma, Sheng-feng Qin, Yong-Jin Liu, *Senior Member, IEEE*, and Hongan Wang, *Member, IEEE*

Abstract— As communications are increasingly taking place virtually, the ability to present well online is becoming an indispensable skill. Online speakers are facing unique challenges in engaging with remote audiences. However, there has been a lack of evidence-based analytical systems for people to comprehensively evaluate online speeches and further discover possibilities for improvement. This paper introduces SpeechMirror, a visual analytics system facilitating reflection on a speech based on insights from a collection of online speeches. The system estimates the impact of different speech techniques on effectiveness and applies them to a speech to give users awareness of the performance of speech techniques. A similarity recommendation approach based on speech factors or script content supports guided exploration to expand knowledge of presentation evidence and accelerate the discovery of speech delivery possibilities. SpeechMirror provides intuitive visualizations and interactions for users to understand speech factors. Among them, SpeechTwin, a novel multimodal visual summary of speech, supports rapid understanding of critical speech factors and comparison of different speech samples, and SpeechPlayer augments the speech video by integrating visualization of the speaker's body language with interaction, for focused analysis. The system utilizes visualizations suited to the distinct nature of different speech factors for user comprehension. The proposed system and visualization techniques were evaluated with domain experts and amateurs, demonstrating usability for users with low visualization literacy and its efficacy in assisting users to develop insights for potential improvement.

Index Terms—Visual Analytics, Multimodal Analysis, Public Speaking, Online Presentation

1 Introduction

In recent years there is a rapidly growing trend of virtual communication. For speakers, the transition from offline to online opens up new ways to express ideas. However, it also poses challenges for speakers to engage audiences at a distance.

Guidance for better virtual presentations has been provided for general tips [47], workplace meetings [42], and various virtual scenarios in remote education [20]. However, these tips only give theoretical advice that can be difficult to be used to evaluate or guide practical presentation delivery. For speaking there are inconsistencies between different theories about speech techniques [32,63]. It is time consuming and experience demanding for speakers to understand how their speech performs and even harder to find references for potential improvement.

Existing literature has reported research on the analysis of speech presentation techniques. Some studies have analyzed the relationship between individual [27,55,56,58,69] or multiple [15,31,37,44,50,62] speech factors and the effectiveness of the speech. However, while these methods typically analyze effectiveness of speech factors in the form of scores, it is challenging for users to grasp the underlying reasons for the score and consequently identify areas for improvement. Visual analytics systems have been proposed to facilitate an interactive exploration of presentation techniques. Existing visual analytics methods can be classified into two dimensions: single factor versus multiple factors, and individual speech versus a collection of speeches. However, these works are unable to directly assess the effectiveness of speech factors, nor do they provide a comprehensive list of speech factors for analysis. Although existing works mainly rely on video inputs, none of them are specifically designed for online public speaking scenarios. Connecting with remote audiences through a camera requires some particular presentation techniques that deserve further exploration.

In this work, we propose *SpeechMirror*, a visual analytics system that allows experts and amateurs in public speaking to gain insights into a speech driven by a large-scale analysis of online speeches. *SpeechMirror* can help understand areas for improvement and recommend examples of speeches as references for practicing improvement.

The system allows enhanced ability to understand the estimated effectiveness of different techniques in a speech. While various ideas exist for measuring speech effectiveness [32], we utilize a collection of videos ranked in a speech contest as our quantifiable metric. To understand the underlying techniques that influence the effectiveness of online speeches, we establish various multimodal speech factors from domain interviews and existing literature. We especially consider techniques that are different in online speeches including the use of stage, eye contact, and body gestures. Based on our collection of speeches, we determine the relationship between various techniques and effectiveness. With these identified relationships, they can be applied to detect the areas for improvements (via diagnostics) and recommend some techniques via examples to improve the speech effectiveness (via prediction). Thus, users can gain awareness of the estimated impact of various factors on a speech as a whole or in individual sentences.

SpeechMirror offers a personalized recommendation approach to explore various possibilities for speech delivery. The recommendation is based on a selected part of speech and produces results at different granularity levels (in entire speech or individual sentence) and different modes (by factors or script contents). The recommendation allows reflection on a speech compared with other speeches in the collection, which expands user knowledge of possible expressions.

Challenges to the scope of the interface of our system include the complexity brought in by a number of multimodal factors, a large amount of changes of the factors over time, as well as supporting

- *Z.Y. Huang, Q. He, X.M. Deng, C.X. Ma, and H.A. Wang are with the Beijing Key Lab of Human-Computer Interaction, Institute of Software, Chinese Academy of Sciences and also with the School of Computer Science and Technology, University of Chinese Academy of Sciences. E-mail: {zeyuan2020, heqiang2022, xiaoming, cuixia, hongan}@iscas.ac.cn.*
- *Kevin Maher is with Diatom Design Limited Liability Company. E-mail: kevintmaher@gmail.com.*
- *Y.-K. Lai is with the Department of Computer Science and Informatics, Cardiff University. E-mail: LaiY4@cardiff.ac.uk.*
- *S. Qing is with the School of Design, Northumbria University. E-mail: sheng-feng.qin@northumbria.ac.uk.*
- *Y.-J. Liu is with the Department of Computer Science and Technology, MOE-Key Laboratory of Pervasive Computing, Tsinghua University. E-mail: liuyongjin@tsinghua.edu.cn.*
- *C.X. Ma, Y.-J. Liu, H.A. Wang are corresponding authors.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

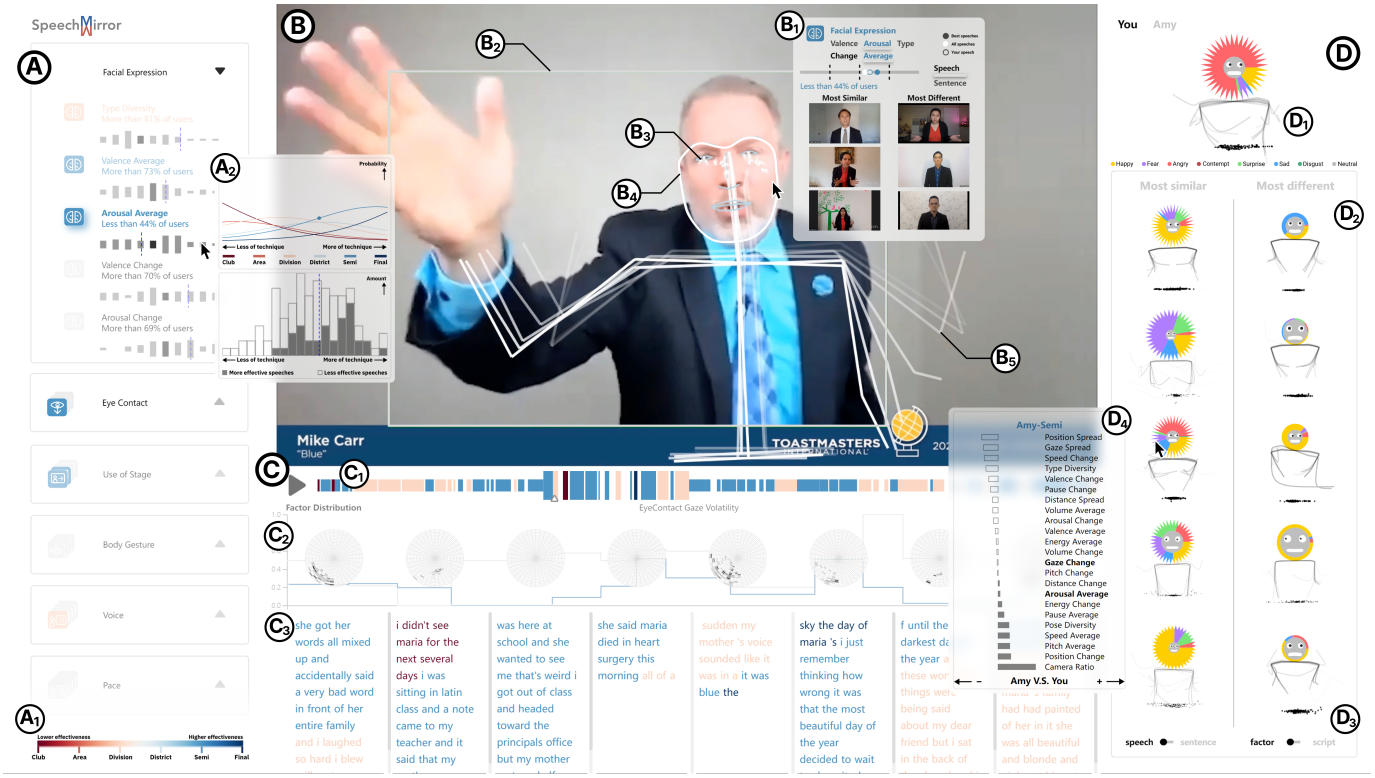

Fig. 1: Our visual analytics system supports the evaluation and understanding of presentation techniques as well as discovering expression possibilities. The Factor Panel (A) delivers feedback on the effectiveness of speech factors and assists in comprehending the trends of speech factor effectiveness. The Speaker Panel (B) provides an augmented video integrating visualizations of presentation techniques. The Time Slice Panel (C) enables users to understand the effectiveness and factor data with script over time. The Mirror Panel (D) recommends similar or different speeches and facilitates intuitive comparison between speeches with visual summary.

comparison and contrast of different speech samples. Given the low visualization literacy of our target users, we provide intuitive visualizations for better understanding. Specifically, we design SpeechTwin, a novel visual speech summary, to facilitate rapid comprehension of speech factors and comparison between speeches. In order to allow an enhanced understanding of body language in the original context, SpeechPlayer directly visualizes presentation techniques in the video feed. Also significant is a series of different glyphs developed to support an intuitive understanding of speech features.

The major contributions of our work include: (1) a visual analytics system to assist in analyzing the effectiveness of an individual online speech within a collection of speeches; (2) novel visualization designs integrating multimodal features and factors to support the understanding of the use of presentation techniques; (3) usability of our system demonstrated and analytical insights gained in an evaluation study.

2 Related Work

2.1 Automated Analysis of Public Speaking

With the assistance of computers, presentations are digitalized and analyzed in a quantitative way. The research can be classified into two major directions in this discussion.

Computation-centered approaches focus on building computational models to evaluate the effectiveness and significance of different presentation techniques. Some research works focused on analyzing presentation techniques in single modality, such as upper body gesture [69], language characteristics [27], body movements [55], prosodic voice characteristics [58], narrative trajectories [56], etc. Additional efforts have been made to take advantage of multimodal information for analysis. To estimate the presentation performance level, Echeverría et al. [15] used eye contact from video, body posture, and body language from a Kinect, and Luzardo et al. [31] used slides and audio. Wörtwein et al. [62] assessed public speaking performance with audiovisual features. Nojavanasghari et al. [37] studied persuasiveness prediction by a deep multimodal fusion method with visual, acoustic, and text descriptors. Ramanarayanan et al. [44] scored presentations with speech, face, emotion and body movement. Sharma et al. [50] predicted video popularity of TED (technology, entertainment and design) talks from facial and physical appearance, facial expressions, and pose variations. These efforts evaluate speech performance based on indicators of effectiveness such as video popularity and audience ratings. In our work, we adopted a speech contest scenario, with the contest placement serving as a measure of speech effectiveness. With a computation-based approach, it is easy for speakers to obtain performance assessment, but hard to further investigate why and how to do better. In our system, we provide insights on speech factor effectiveness and further support exploration for areas of improvement.

Visualization-centered approaches present the data of presentation techniques through visualizations and further support users in gaining insights through exploration. Existing literature focuses on different purposes of speech analysis: identifying narration strategies [64], understanding emotion [67], training vocal skills [60], analyzing debate transcripts [52], exploring presentation techniques [63], decomposing humor [59], assisting the exploration of gestures [66], and analyzing the effectiveness of different speech factors [32]. While prior works have provided approaches to understand the contributions of speech factors or recommend evidence for public speaking training, our work aims at evaluating a single speech, assisting users to analyze the speech in the context of a speech collection.

Current research mainly focuses on offline presentations or taking a video as input. Distinct from in-person presentations, there are unique techniques that can make online presentations more effective. To the best of our knowledge, there is still a lack of qualitative research in the online public speaking scenario.

2.2 Online Public Speaking

The importance of online public speaking has increased in recent years. The consulting firm Gartner predicts that by 2024 only 25% of business meetings are offline [46]. In the testing phase of a system including both online and offline speeches, an experienced public speaking expert

reasoned that there were significant differences in the emotional expression of online and offline contests [32]. Teodorescu et al. [57] found possible differences in NLP (Natural Language Processing) of titles in an online and offline contest, focused on microcontroller applications.

There are other important differences of online and offline speeches that could be uniquely investigated by analytical systems. Since the fixed reference frame of the camera and fixed position of the microphone are perceived the same to all members of the audience, the effects of eye contact, stage movement, and voice volume can be better measured. Ochoa et al. created an automated system measuring speech delivery to a remote audience [38].

2.3 Human Physical Behavior Data Visualization

Our work focuses on body language, including facial expression, eye gaze, gestures, and positional movement.

Gesture and positional information is important to understand in speeches. In sports, Stein et al. [53] integrated movement data including possible player movement directly with the original video, "enabling analysts to draw on the advantages of both". Mova [1] used small multiples of extracted body keypoints that were annotated with color to indicate features of movement. A design prototype we considered included these enhanced keypoints. In public speaking, GestureLens [66] provided interactions to support the understanding of gestures in relation to content and time. In contrast to their work, we focus on how to show gestures not only between different words in a sentence, but also creating representative skeletons for custom time ranges.

There are many strategies to visualize eye gaze. Much research focused on how to visualize where gaze is directed, such as fixation points and saccades [13], and heatmaps [39]. Our work instead focuses on the direction that a speaker is looking at, which can more clearly show eye contact or gaze directions that fit with speech content. Similarly, Higuch et al. developed visualizations to show the eye gaze direction of autistic children [22].

The complexity of facial expression has led to a variety of visualization strategies to present them. E-ffective used a spiral visualization to show emotional shifts in speakers as well as a text-based visualization to display emotion in the context of a speech script [32]. In EmotionCues, Zeng et al. created an Emotion Band, a visualization that allows users to track emotion changes of multiple users in a flow-based design [65]. For our work we were interested in presenting the relative differences for a range of emotional factors, and developed visualizations for a range of these factors.

3 Design of SpeechMirror

In this section, we will introduce the domain-centered design of *SpeechMirror*. The design process mainly contains two stages: literature reviews and domain interviews with public speaking experts and amateurs. With the insights gained, we derived the design considerations and the scope of presentation techniques in our system.

3.1 Design Process

The main goal of our work is to assist public speaking experts and amateurs in understanding a speech by means of guided insights from a collection of speeches. We first reviewed the existing literature to build a list of speech factors that are potentially related to the effectiveness of public speaking. We also developed an understanding of the distinct characteristics of online speech presentations.

To better understand the demands amateurs and experts valued as important, we conducted 30-minute semi-structured interviews with 6 volunteer participants, including 4 experts (DE1 - DE4) and 2 amateurs (DA1 - DA2). Both the amateurs and the experts had participated in the World Championship of Public Speaking (WCPS) contest at least once. The experts all had experience coaching public speaking as an occupation. The amateurs have experience in online public speaking contests between 2-5 times and have watched no more than 50 speeches.

The goal of the interviews is to understand their opinions on effective online presentation techniques, current practice and challenges in developing public speaking skills, as well as their needs for a new speech video analysis tool. Detailed information on the interviews is provided in Sect. 1 of the supplemental material.

3.2 Design Considerations

We derived the design considerations for *SpeechMirror* from existing literature and our domain interviews.

DC1: Provide an integrated way of understanding the effectiveness of speech factors in a speech within a speech collection. It is important to reveal the effectiveness of speeches. In our interviews, DE4 mentioned the requirement of having a benchmark to measure speech performance. Focusing on providing exploration of factors within a speech collection, E-ffective showed both the importance and difficulty in understanding the effectiveness of speech factors through the feedback ratings. To migrate the challenges, we sought to provide users' insights on their speeches in *SpeechMirror*, so that users can understand the speech performance. In the interviews, DE3, a full-time public speaking trainer, thought there are many factors in speeches that trainers need to keep in mind for feedback. He remarked, "*a trainer might ignore or forget the top 8 (factors), but a system might be able to tell the top 9.*" Thus we sought out a more comprehensive set of effectiveness factors for consideration, as well as a more integrated way for understanding the factors.

DC2: Provide a convenient approach for obtaining speeches for reference. A promising study [60] demonstrated the potential for improvement of speeches by discovery of high quality examples. They claimed the future improvement of showing negative samples as warnings to avoid. The importance of references was also brought to our attention during our interview sessions. In our interviews with less experienced amateurs, they voiced needs to discover different ways of expression. When discussing the use of body language, DA2 stated "*a problem is that I only compare to what I know. With the same (speech) content, I wonder how others express it.*" DA1 similarly expressed uncertainty about the use of eye contact and gestures. More generally, DE2 claimed priority should be placed in video referencing, "*study the people who do better than you are.*" Our system aims to offer speech samples with similar or different speech delivery or content to facilitate exploration. These samples exhibit varying levels of effectiveness to allow the comprehension of possible expressions.

DC3: Reveal the temporal distribution of the effectiveness of speech factors. The order of techniques in speeches matters, and many sources have theories about advantages of using techniques at different times. The WCPS 2012 champion advised stage movement where the speaker should "end your speech in the same location where you began" [14]. A university textbook on public speaking claims the order of gestures as they appear in relation to the main idea in the speech can be important [33]. Our interviews also revealed opportunity for speakers to learn from time order. DE3 claimed speakers might not only be unclear about what they should do in their speech, but also "*not be clear what they did in their speech. Did they smile at the beginning?*" *SpeechMirror* supports the needs of understanding the use of presentation techniques over time in an intuitive way.

DC4: Demonstrate speech factors in relation to the multimodal context within a speech. Our interviews revealed the necessity of viewing speech factors, such as the factor raw data and factor effectiveness, in reference to the broader context of the speech. In our interview DE4 stated that "*content is the most important, especially cultural background*", pointing out the significance of verbal context. Displaying how non-verbal techniques relate to the verbal context is commonly used in visual analytic systems for speech analysis, e.g. [32,59,63,66,67]. The video context was also claimed to be important in our interviews. DE3, DE2, DA1 stated that for online speeches, understanding how to move in the limited space was crucial. Other interviewees brought up factors such as eye contact, stage movement, and using gestures within the limited space, all of which are difficult to understand without reference to the original video. Thus, we intend to provide the original video context and verbal context to enhance understanding of factors inside a speech.

DC5: Summarize the most relevant techniques used. We evaluate 23 different techniques in speeches. Given the large amount of informa-

Table 1: Presentation techniques, multimodal features and corresponding factors considered in *SpeechMirror*. The significance of factors with * indicates a significant correlation with speech effectiveness ($p < 0.05$).

Presentation Technique	Feature	Factor	Significance
Facial Expression	Type	Diversity	$p = 0.002^*$
	Valence	Volatility	$p = 0.571$
		Average	$p = 0.005^*$
	Arousal	Volatility	$p = 0.761$
		Average	$p = 0.000^*$
Eye Contact	Gaze Direction	Volatility	$p = 0.002^*$
		Dispersion	$p = 0.067$
	Watching Camera	Ratio	$p = 0.265$
Use of Stage	Distance from Camera	Volatility	$p = 0.908$
		Dispersion	$p = 0.185$
	Position in Frame	Volatility	$p = 0.026^*$
		Dispersion	$p = 0.141$
Body Gesture	Gesture Energy	Volatility	$p = 0.860$
		Average	$p = 0.426$
	Gesture Diversity	Diversity	$p = 0.266$
Voice	Volume	Volatility	$p = 0.000^*$
		Average	$p = 0.413$
	Pitch	Volatility	$p = 0.438$
		Average	$p = 0.988$
Pace	Speaking Rate	Volatility	$p = 0.617$
		Average	$p = 0.198$
	Pauses	Volatility	$p = 0.157$
		Average	$p = 0.533$
Content	Script	-	-

tion at hand, visualization methods that provide relevant summaries of the techniques are crucial. The data of the techniques are time series, so it is hard for users to grab key information within the data. Data aggregation and abstraction are thus necessary. Relevancy for the summaries both depends on the significance of the factor and the different use of the technique in terms of speeches in the collection. Less significant information is deemphasized or hidden from view. The summaries aim to enhance browsing of speeches, provide key information about the user's speech, and assist comparison of different speeches.

3.3 Scope of Presentation Techniques

There are many techniques that go into successful public speaking. Schneider et al. [49] provide an extensive list collected from public speaking experts about non-verbal communication practices.

For online speaking there are additional techniques to consider, due to the limited camera view, single viewing angle of the audience, etc. Several techniques critical to effective online speaking have yet to be intentionally visualized for further analysis. Through existing literature and domain interviews, we attempt to build a more comprehensive list of techniques (**DC1**). We balanced the technical feasibility and the domain significance in our list.

Eye contact is a technique we found important enough to focus on. DE2 claimed that in online contests "*connection to the audience with a camera is critical. The key is to be a natural speaker in front of the camera.*" DE1 and DA1 also found its importance, with DA1 claiming they "*don't know where to look at.*". Literature on online public speaking also emphasized the importance of eye contact. In the book *Presenting Virtually* [47], Patti Sanchez claimed "eye contact is essential to make your message feel direct and personal", however, "looking directly into the camera can be a challenge for presenters who are accustomed to speaking in person". Data from online speeches provides the ability to assess the effective use of eye gaze, since there is a universal viewing angle of the speech by all online members.

Body gesture and use of stage are important for online speeches. In open discussion, four of the six interviewees thought the impact of the screen in online interactions for body language was especially important. However, some of them also mentioned the difficulty of keeping various motions within the limited space of screen. This also reflects the importance of the proper use of "stage" in the camera. DE2 questioned: "*You are the director with a fixed camera. Do you need to show your full body above your waist?*" Online speaking enables accurate calculation of stage and gesture occlusion by providing a consistent camera angle, unlike offline speech where the audience's viewing angles vary.

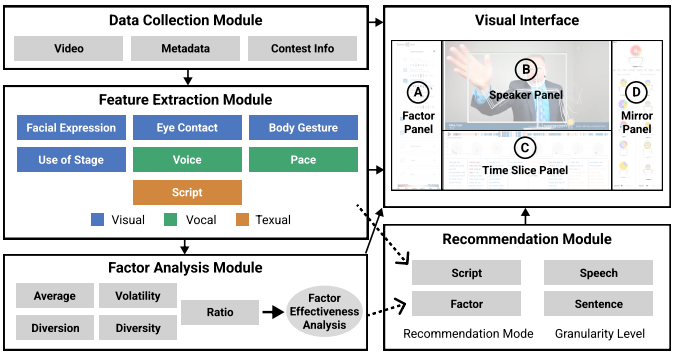

Fig. 2: System architecture of *SpeechMirror*.

Emotion plays an important role in public speaking. In our interview survey, emotion related factors ranked high in the list while emotional diversity was ranked the most important factor by interviewees. Emotion can be explicitly expressed by the speakers' facial expression, voice and text content. In our review of several methods of extracting vocal and textual emotion, we found them less accurate than the facial expression results, even with the state-of-the-art models. Therefore, this work mainly focuses on facial expressions.

Vocal characteristics include voice volume, pitch, speaking rate, and pauses. Proper speaking rate and pauses are difficult because there is no feedback from the audience in the online contest. DA1 stated "*Faces of the audience are not shown, and no audio reply is allowed. You don't know if the audience gets your points or joke, and can't change without feedback.*" While volume and pitch were not mentioned by the interviewees, they are commonly considered important in public speaking literature. Online speakers are required to have good volume management skills to control their speaking volume.

4 System and Data

4.1 System Overview

We design and implement a visual analytic system, *SpeechMirror*, to fulfill the analytical goal and the design considerations. Our system integrates raw video data, extracted multimodal feature data, speech factors and corresponding effectiveness data to provide a complete data processing and analyzing workflow. All data except the raw video is stored in MongoDB for fast access.

As illustrated in Figure 2, our system consists of five major modules. The *Data Collection Module* contains the speech videos and information we collected for analysis. The newly input video will also be stored in the module. The *Feature Extraction Module* extracts features from the input video, including visual, vocal and textual modalities. The *Factor Analysis Module* determines speech factors based on the extracted features and further estimates the effectiveness of factors. The *Recommendation Module* searches for the most similar and different speeches from the video collection. The *Visual Interface* provides visualizations of data and interactions to support analysis. The interface is designed and implemented in a browser-server architecture, utilizing d3.js [8] for creating visualizations in the front-end interface, and Flask framework [40] for providing web services on the back-end.

4.2 Data Collection Module

We manually collected a collection of speech videos in the World Championship of Public Speaking (WCPS) contest from public online platforms like YouTube. The entire online speech video collection in our system contains 102 videos in total. Each speech is about 7 minutes long and of good visual-audio quality. We recorded the metadata (including the start and end of each speech in the video), and contest information (including region, year, level, and rank). Whether the video is delivered online or offline was also labeled.

4.3 Feature Extraction Module

To assist users with quantitative analysis of presentation techniques, the feature extraction module takes video as input and automatically extracts and processes features for further analysis. The module extracts

the following speech-related features from multimodal inputs including visual, vocal, and textual. The multimodal features are aligned based on the timestamps of the script's words.

- **Facial expression:** The face of speaker in each frame is detected by face_recognition [18] and clustered out from other faces by DBSCAN [16]. We apply AffectNet [35] to predict valence and arousal values (ranging from [-1, 1]) from face images. AffectNet is a widely used baseline method for predicting facial expression, valence and arousal from images in the wild. A convolutional neural network [2] is used to classify face images into seven classic emotion categories [19].
- **Eye contact:** We apply OpenFace Toolkit [3,61] to estimate the eye gaze direction of both eyes. The eye gaze direction of each eye is represented as a normalized 3D vector in world coordinates. Average eye gaze direction is converted to radians in world coordinates. The angle of watching the camera is calculated by averaging the angle obtained from the 3D position vector of the eyes relative to the camera and the vector of the eye gaze direction.
- **Body gesture:** We adopt MMPose [12], a widely used open-source toolbox for pose estimation, to predict the 2D body keypoints from videos. We use Faster R-CNN model [45] with a ResNet-50-FPN backbone for human bounding box detection. HRNet [54] pre-trained on COCO [30] detects 2D keypoints from video frames with the Human3.6M [23] format. We set up rules to retain the body keypoints of the speaker. We calculate the kinetic energy [36] of the speaker's upper body as the gesture energy based on the offset of the keypoints in adjacent frames and the mass distribution of human body [41]. Gesture diversity is determined by the standard deviation of the cosine distances between the upper body keypoints of each frame and the first frame. The body keypoints are first aligned by the thorax and normalized to a fixed width of shoulder.
- **Use of stage:** We utilize the z-axis position of the speaker's head relative to the camera estimated by OpenFace Toolkit [3], as the distance between the speaker and the camera. The position of the speaker in the video frame is determined by the center of the bounding box detected by MMPose during body gesture extraction.
- **Voice:** The loudness and pitch data is extracted by Parselmouth [24], a Python library for Praat [7] which is a widely used speech analysis software in phonetics. We use the "sound to intensity" and "sound to pitch" functions to capture the loudness (in dB) and frequency (in Hz) of a speech, respectively.
- **Pace:** We compute the pauses between different words and different sentences according to the timestamps of each sentence and word. The speaking rate is determined by the duration per syllable of each word. The syllable of word is counted by the vowel sounds in a word with the CMUdict corpus [26] in NLTK Language Toolkit [5].
- **Script:** Each video is transcribed by Azure Cognitive Speech to Text Service [34] with timestamps of each sentence and word. We use the Universal Sentence Encoder [10] to encode script texts into 512 dimensional vectors while maintaining the semantic information.

4.4 Factor Analysis Module

The visual and vocal features extracted from the videos are time series, which are difficult for understanding and comparison. Considering the domain requirements, we calculate the factors as shown in Table 1.

The methods of factor calculation are as follows: **Average** represents the mean value of data over time. **Volatility**, representing data change over time, is calculated using the CID algorithm [4] with normalization. CID measures the complexity of time-series data, capturing patterns like peaks and valleys. **Dispersion** is determined by the variance of the time-series data, obtained by dividing the standard deviation by the mean. **Ratio of watching camera** indicates the proportion of frames that the speaker is looking directly at the camera within a 5-degree angle. **Diversity of facial expression type** represents the variety and relative abundance of the emotions [43]. The time-series emotion type data are represented as $D = \{d_t\}_{t=1}^{T}$, where d_t indicates the t-th sample of emotion type data. Let r_i denote the proportion of the same emotion type as d_i in D. Diversity is calculated in $diversity = \sum_{i=1}^{T}(r_i \times \ln r_i)$.

With the speech factors and contest placements of speeches indicating speech effectiveness, we build the effectiveness relationship of each factor with a regression method [21]. Specifically, we conducted parallel line tests and observed significance with $p < 0.05$. Then a multi-class ordinal regression is employed to evaluate the significance of speech factor effectiveness and the regression results are presented in Table 1. We predict the effectiveness of each factor based on its significant relationship with speech effectiveness.

4.5 Recommendation Module

To enable users to quickly find the presentation examples that are similar or different to the speech for analysis (**DC2**), our system recommends relevant speeches from our data collection for their reference.

The recommendation module grants users the option to manually select from **two granularity levels** (speech or sentence) and **two recommendation modes** (factor or script) through the interface. The granularity levels determine the extent of the search range, which spans either the entire speeches or individual sentences. The recommendation modes dictate the method of similarity calculation, which can either rely on speech factors or the transcribed script content.

The recommendation consists of three steps: (1) Prepare the query and candidate data. The module calculates the query data of selected period in the analyzed speech, as well as the data of candidates according to the selected granularity level. (2) Extract the feature vectors for both query and candidates. For the factor recommendation mode, we join the values of the factors selected on the interface into a vector after a min-max normalization. For the script recommendation mode, we use the textual semantic embedding vectors for both input and candidate scripts. (3) Fetch the most similar or different candidates to the query. We calculate the similarity distances between the query and the candidates by Euclidean distance for vectors in the factor mode and cosine distance for vectors in the script mode. Heap queue algorithm is used to obtain the results with largest or smallest similarity.

5 User Interface Design

In this section, we introduce the general principles for design, the interface, and visualization designs.

The design of our visualizations aimed to be intuitively understood by an audience with minimal visualization literacy, which would allow our interface to be used by a wider audience with minimal training. We iterated our design closely with potential users and in the process created design principles that guided our work. In our initial designs we found significant difficulty in understanding the concepts in our interface. In order to increase the intuitiveness of our system, we sought design principles that followed several principles given by Blair et al. [6] and reflections by Böttinger et al. [9], including: **(DP1)** making consistent use of visual elements, **(DP2)** providing elements in proximity to the content, **(DP3)** offering interactions as direct with content as possible, **(DP4)** showing understandable explanatory visualizations to a broader audience. The design alternatives in our design iteration process are introduced in Section. 4.1 of the supplemental material.

5.1 Icons and Color Encodings

Repetition of elements in the design of icons, as shown in Figure 3, was used to both aid in the understanding of intricate concepts and used throughout the three panels in the system to allow rapid understanding. **(DP1)** The icons were designed as pictographs that resemble the concepts they are linked to. For example, the position of a speaker on the screen was given by a figure icon surrounded by a screen.

We applied consistent use of color encoding for effectiveness across the three panels of our interface using a diverging color scale that emphasizes the data of two extremes, as shown in Figure 1 (A1). **(DP1)** This color scheme is a scale with dark red signifying very low effectiveness metrics to a dark blue signifying very high effectiveness metrics. A light gray is used to indicate factors that were not significantly related to effectiveness. Another color encoding scheme for emotion is employed in SpeechTwin, which will be further introduced in subsubsection 5.3.1.

5.2 Interface Panels

Our interface is composed of four panels that support the design considerations as described in subsection 3.2.

Fig. 3: Consistently used icons for the presentation techniques.

The Factor Panel presents an overall view of factors that appear in a speech, and how they relate to all speeches in our video collection (**DC1**), as shown in Figure 1-A. The 23 speech factors considered in our study are categorized into 6 groups of presentation techniques. The panel provides a summary of the factors when a group is closed, and more detailed information when opened. Each factor is represented by colored icons and text that indicate its effectiveness, as explained in subsection 5.1. **(DP4)** When a user hovers over a factor, the factor effectiveness board (A2) will be displayed, showing the effectiveness trend and the factor distribution in the video collection. Factors selected in this panel will filter the results in the other three panels. If no factor is selected, then the aggregated results of all factors are shown.

The Speaker Panel supports understanding speaking techniques with direct reference to the original context in the video. The techniques are visually presented, overlaid on the video with reference to the regions they correspond to. Users can directly interact with the layered visualizations to focus on a specific technique, thereby displaying the presentation technique board (B1). Data shown in the panel comes from the current playing sentence. Once a factor is focused, more detailed information about the use of the factor is shown, as well as a recommendation of speeches for reference (**DC2**). Users are allowed to play the recommended videos and set one to the comparison video. The comparison video will be displayed at the top-right corner of the Speaker Panel and can be switched to the main video in the panel.

The Time Slice Panel enables users to select different parts of a speech, as well as provides different views of speech data with script context over time (**DC3, DC4**), as shown in Figure 1-C. At the top of the panel, the timeline (C1) shows the use of selected factors over time. Each rectangle represents a sentence, with the color encoded as described in subsection 5.1. White blanks indicate the intervals between sentences. A brush tool is provided for convenient selection of speech periods on the timeline. The selected period is divided into 8 time slices, using glyphs (C2) to show more detailed use of the factor and raw data. At the panel bottom is a text module (C3) highlighting the effectiveness of factors in reference to the speech script. Words that coincide with the time slice segmentation points are split into two parts based on time, allowing for a better comprehension of the speaking pace within shorter selected time periods.

The Mirror Panel allows speakers to find the most similar and most different speeches compared to the whole or a part of a speech (**DC2**), as shown in Figure 1-D. At the top of the panel, a visual summary of speech factors (D1) for the selected part of the speech, named SpeechTwin, is presented to facilitate quick comprehension of speech factors. Below, the panel demonstrates the SpeechTwins of similar and different speeches (D2) according to the selected recommendation mode and granularity level (D3). Hovering over a SpeechTwin will trigger the speech comparison board (D4) while clicking on it will set the speech as the comparison video. Users are allowed to switch between the original video and the comparison video by clicking the video name located at the panel top.

5.3 Visual Design

5.3.1 SpeechTwin: A Multimodal Speech Summary

Our system provides a visual summary of crucial speech techniques in an intuitive way by mapping the technique to symbols on a figure (**DC5, DP2**). We display the visual encoding used and the demonstration of SpeechTwin in Figure 4. More demonstrations are shown in the supplemental material (Sect. 4.2) and supplemental video for reference.

Facial and vocal data are represented on a Chernoff face [11], which allow these features to be combined and closely associated with corresponding facial regions. Eye gaze direction is communicated by the angle the eyes are looking at (Figure 4 A-V). The positivity or negativity of facial expression, namely valence, is given by the upward or downward turn of the mouth as a smile or frown (A-III). The intensity

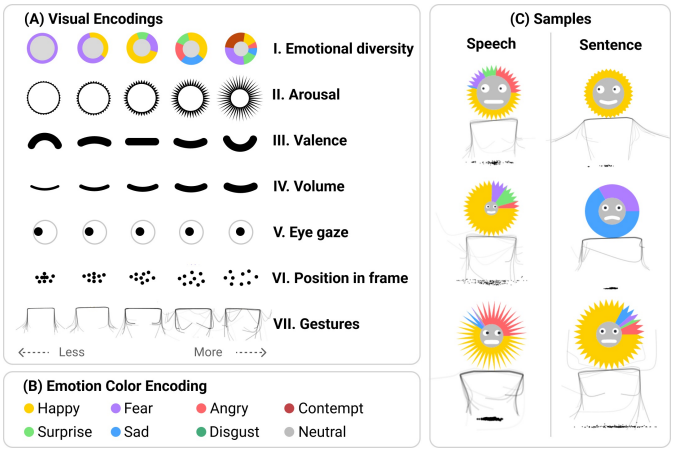

Fig. 4: SpeechTwin: a novel visualization of multimodal speech summary.

of facial expression, arousal, is conveyed by the protrusion of spikes outside the face, as inspired by the coding of arousal in shape [51] (A-II). The emotional diversity is indicated by the area of face given to emotions (A-I), with a neutral gray emotion in the center, and other emotions shown in the area of the outside circle and with different colors displaying different emotions (Figure 4 B). The loudness of voice is represented by the width of the mouth (A-IV). The eyes and the mouth maintain a fixed proportion in relation to the face. To ensure the readability of the elements inside the face in rare cases of scarce neutral emotion, we set a minimum size to the face.

To better describe various gestures in the speech while minimizing visual distractions, the representative body gestures are given in the arms and shoulders of SpeechTwin. The upper body keypoints of the speaker in each frame are aligned by the position of the thorax, and normalized to a fixed shoulder width. Inspired by PoseTrans [25], we use a Gaussian Mixture Model (GMM) to classify the poses into 10 clusters. We identify the pose with the smallest sum of cosine distances to all other poses in the cluster as the most representative one. We visualize the most representative gestures with the opacity indicating the amount of gestures in the corresponding cluster (A-VII).

The use of the stage is described by the "footprints" or dots beneath the character, with each dot representing the center of the speaker's bounding box on the screen (A-VI).

Hovering over a SpeechTwin in the Mirror Panel (Figure 1-D2) will trigger the speech comparison board (D4). On the board all factors of the compared speaker are contrasted with the original speaker. Factor differences between the speeches are shown on a bar chart, with the higher differences polarized at the top and the bottom. The placement of the speech in the contest is also shown.

5.3.2 SpeechPlayer: An Augmented Speech Video Player

Considering the potential information omission when viewing speech videos, we offer SpeechPlayer (Figure 1-B), an interactive approach to enhancing users' understanding of techniques during video playback, as well as augmenting their sense of expressive possibilities (**DC4**).

SpeechPlayer integrates the visualization elements of critical presentation techniques directly in the video feed. **(DP2, DP3)** The facemark keypoints are displayed on the video frame to emphasize the facial expressions of the speaker (Figure 1-B4). The direction of eye gaze is depicted through a ray, which becomes increasingly transparent along the direction of the gaze (B3). We found with traditional methods of an opaque line, confusion about the direction of eye gaze can occur when viewing eye gaze over time, such as when speakers change their head angle to act different characters. The skeleton of the speaker's upper body (B5) and the bounding box of the speaker (B2) are also visualized to enhance the understanding of body gestures and the positions in video frames.

To enhance the understanding of eye gaze and body gestures across time, SpeechPlayer displays 10 skeletons and eye gaze rays at fixed time intervals. The transparency of color corresponds to the temporal

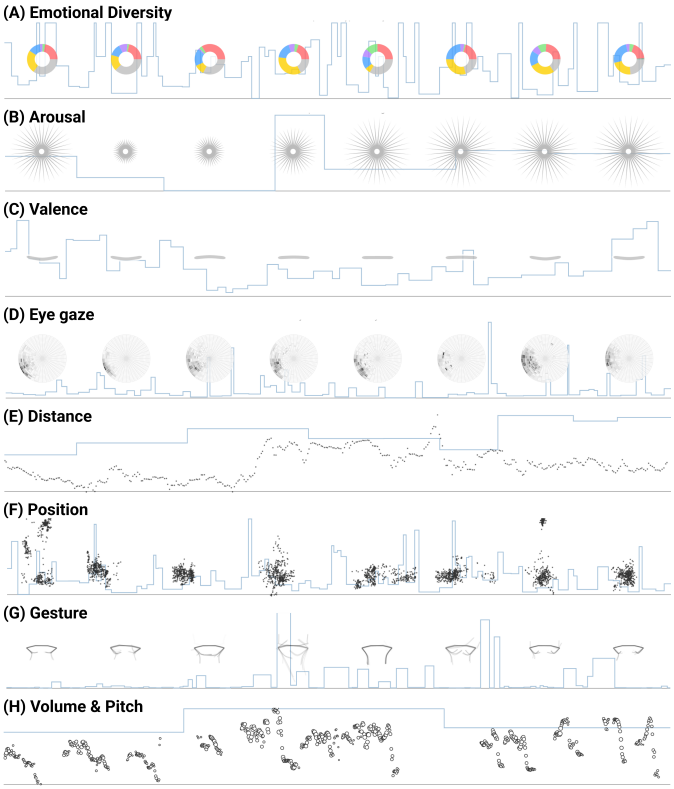

Fig. 5: Demonstration of modality feature visualizations.

proximity to the current playback time, with higher opacity indicating closer proximity to the current time.

Users can interact directly with the displayed elements on SpeechPlayer, and hovering allows analysis of a group of presentation techniques with the presentation technique board (Figure 1-B1). Users can choose a speech factor of interest, and compare the distribution of the factor values of the currently played sentence with the average factor value of best speeches and all speeches (**DC1**). Speeches or sentences that have the most similar and most different use of a factor are recommended so users can rapidly find reference samples (**DC2**).

5.3.3 Visualizations of Multimodal Features

We offer visualizations in the Time Slice Panel to illustrate the temporal details of modality features. These visualizations are tailored to the specific nature of original feature, as shown in Figure 5.

For each feature in facial expressions, we use visualization methods consistent with SpeechTwin, as shown in Figure 5-A, B and C respectively. **(DP1)** In an informative visual summary of gaze patterns, a gaze heatmap is used to depict the frequency of eye gaze directions. Darker shades on the heatmap indicate higher frequencies. To better understand body gestures, we considered several visualization methods as mentioned in subsection 2.3. Balancing between the need for simplicity and complexity of gestures in a large number of video frames, we display representative gestures that provide an uncluttered view of representative gestures, as shown in Figure 5-G. The data in each time slice are aggregated to summarize the features.

We offer two visualizations to depict the use of stage techniques. To understand the speaker's positions in the video frame, we display the centers of speaker's bounding boxes in each time slice within a rectangular area, as shown in Figure 5-F. This allows us to observe the range of speaker's position changes throughout the time slice. Additionally, we apply a scatter plot to illustrate the changes in distance between the speaker and camera over time in the selected speech period.

For the features in the voice technique, we use a scatter plot to visualize volume and pitch data along time referring to the work of Schaefer et al. [48], as illustrated in Figure 5-H. We map the x-axis as the timeline, with the y-axis representing pitch, and the radius of each point indicating volume. Users can observe the changes in volume and pitch over time directly through this chart.

5.4 Interactions

The proposed system, *SpeechMirror*, enhances users' analysis abilities while maintaining intuitiveness and fluency through various ways of interaction within the four linked panels of the system. **(DP3)** Here we summarize the supported interactions:

Clicking. In the Time Slice Panel, clicking a word within the text module will trigger the video in the Speaker Panel to jump to the corresponding start time of the word. Double-clicking the word will select the sentence that contains the word on the timeline. Users are allowed to activate the comparison video in the Speaker Panel by clicking the SpeechTwin in the Mirror Panel. Clicking on the video will play or pause the comparison video, and double-clicking will switch the video in Speaker Panel to the comparison speech, along with changing the corresponding visualizations in the Time Slice Panel.

Selecting. Our system enables users to select any factor or a group of factors in a feature for analysis in the Factor Panel. The selection of a factor will result in the Time Slice Panel switching to the corresponding factor data, and the SpeechTwin of the analyzed video in the Mirror Panel being refreshed, thus generating new recommendations. Users are allowed to select different recommendation modes and different granularity levels with the toggle buttons in the Mirror Panel.

Hovering. Hovering on the histogram in the Factor Panel displays the factor effectiveness board, while hovering over the data view in the Speaker Panel reveals the presentation technique board for more detailed information. In the Time Slice Panel, when users hover over the line chart of factor distribution, the time-aggregated data of the hovered feature will be shown above. Furthermore, when users hover over the SpeechTwin in the Mirror Panel, a new panel will be shown to reveal the difference between factor values of the speakers.

Dragging and Brushing. In the Time Slice Panel, users can drag the white triangle on the timeline to change the play position on the video. Users can brush a selection area on the timeline or drag the selected area to focus on a sequence of sentences. The Time Slice Panel and the Mirror Panel will refresh accordingly.

6 Evaluation

We aim to evaluate the proposed system and visualizations in two scenarios based on the framework introduced by Lam et al. [29]: user experience (UE) and visual data analysis and reasoning (VDAR). We used a user experience questionnaire and a follow-up interview to elicit the subjective feedback and opinions on our system and visualizations (UE). A case study is used to demonstrate how users explore data and find insights with our system (VDAR).

6.1 Study Design and Procedure

A user study was designed to assess the performance of our system and visualizations in the two scenarios.

Since experts and amateurs in public speaking are the target users of our system, we aimed to gain insight into their diverse usage patterns and perceptions of the system. For amateur participants, we requested their recent online public speaking videos for analysis in the evaluation. Expert participants were asked to analyze a speech from our data collection to simulate critiquing a student's speech as a coach.

The procedure of our evaluation study was split into three sessions. The whole study lasted about 90-100 minutes.

Introduction. We first introduced the main goal and background of our system. After obtaining informed consent from the participants and collecting their personal information, we provided a detailed introduction of the important concepts to facilitate their understanding and utilization during subsequent sessions. By means of quick questioning, we confirmed that participants had fully understood the concepts we presented. The introduction session took about 20 minutes.

Exploration. The user study contains two phases for the participants to explore our system: user tasks and free analysis. We guided the users to complete the user tasks while we observed their use of our system and assessed usability. The free analysis phase aims to observe how

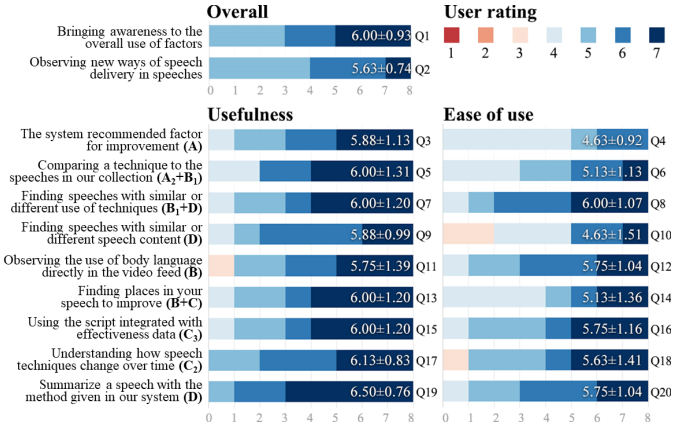

Fig. 6: Questions and results of our user experience questionnaire. The white overlay in the bars indicates the Mean±SD. The target of each question is indicated by the labels in Figure 1.

participants analyze speeches with our system by allowing them to use the system on their own. They were encouraged to think aloud while using the system and voice their findings during the analysis. Both phases took about 30 minutes.

Reflection. Based on their experience with the system, participants were requested to complete a user experience questionnaire, and were encouraged to provide their opinions for each question. We further conducted a follow-up interview to gather additional feedback and insights about our system and its potential actual applications. This session took about 10 minutes.

6.2 Participants

We recruited 8 participants for the evaluation: four experts (EE1 - EE4) and four amateurs (EA1 - EA4) in public speaking. None of the participants participated in the previous domain interviews. The experts all have more than 3 years in coaching for public speaking and have participated in the contest. The amateurs did not coach public speaking as an occupation, and all had participated in the contest. The participants were compensated for $94 per expert user and $54 per amateur user. Detailed information of the participants in the user study is provided in the Sect. 5.1 of the supplemental material.

6.3 User Experience Questionnaire

A user experience questionnaire was employed to assess the participants' subjective feedback of our system and visualization approaches (UE). We adopted a 7-point Likert scale to score the experiences of participants. The questions with their evaluation target and corresponding scores are illustrated in Figure 6. We present the ratings of all participants for each question in the supplemental material.

In general, participants appreciated that our system helped them become aware of the usage of presentation techniques in speeches (Q1) and discover new ways of expression (Q2). On both of these questions, our system received an average score of 5.81 out of 7, which indicates the participants' overall acceptance of *SpeechMirror*. For the usefulness and user-friendliness of use of different parts of our system (Q3-Q20), we received an average score of 5.69, indicating the participants' general approval of the methods we adopted.

The participants acknowledged the usefulness of our system, with an average score of 6.01/7 for usability. Specifically, the participants were satisfied with the usefulness of viewing techniques over time (Q17 for the Time Slice Panel and the visualizations of modality features) and the summary of presentation techniques (Q19 for the SpeechTwin). One amateur participant rated 3/7 on the usefulness of the Speaker Panel and SpeechPlayer (Q11), while other participants provided positive ratings. She suggested providing an option to enhance the original video with an on/off toggle or in a separate window.

The participants' satisfaction with the ease of use was lower, with an average score of 5.38. There are two questions in our questionnaire that were rated below 5.0 on average. The concept of factor effectiveness and related charts required more effort to understand for participants, especially those without a STEM background. Additionally, providing users with multiple factors could also result in cognitive burden, as users reported feeling unsure about where to start exploring factors. (Q4 for the Factor Panel) Recommendation based on speech content was rated as 4.63. Several participants found the recommendation results inaccurate. This is likely due to our limited database and lack of samples for comparison. (Q10 for the Mirror Panel)

6.4 Follow-up Interview

We conducted a semi-structured interview to gather further feedback and insights on *SpeechMirror* from the participants (UE).

Factor Analysis. Our system received positive feedback on providing nearly comprehensive speech effectiveness analysis. EA3 said "*It is amazing to be aware of so many factors. I didn't imagine this was possible.*" EE3 commented that "*all these factors are very useful*". EE4 considered the analysis of our system to be very interesting. After watching the whole speech of his "student", he confirmed that the result of our system matched his judgment of the speech as an expert coach. However, EA1 mentioned feeling information overload, feeling "*overwhelmed due to a lot of factors*".

Speech Recommendation. The participants appreciated the recommendation function since it took less effort to find other speeches for comparison. EA2 said that "*I have never thought about who is the most similar or different to myself. Before, it was hard for me to find someone to compare to.*" EE1 and EE2 both mentioned that the SpeechTwin demonstrated the "speech style", allowing users to compare it with others and "*find a personalized speech style*". EE2 and EA1 stated the benefit of the Mirror Panel and SpeechTwin in enabling users to "*find a comparison and reference*" and "*know who to learn from*".

System Complexity. The participants indicated a preference for a simplified system to improve user experience, suggesting that the current system is overly complex. EE1 reported a high learning curve and instead wished for more "*visual reminders for key functions and terms.*" EA2 and EA3, with non-STEM backgrounds, also suggested adding tooltips in the system for better understanding. EE4 suggested a potential simplification of the interface by "*filtering the most important elements*". EA2 and EE4 observed that the text script within the Time Slice Panel was not very legible, potentially because there were too few words on each line in each slice.

Potential Application. In the interview, we gathered valuable insights from the participants regarding the potential applications of our system. We also observed a notable level of anticipation for further utilization of the system. EE3 stated that she planned to have students review the system-generated results and analyze presentation techniques with the system in tandem with her class material. EA3 noted "*I want to put all my speech videos into the system and see where to improve. The system directly tells me how differently others present, so I can learn from others*". We see a promising application in the future of our system or other similar systems.

6.5 Case Study

During the free exploration phase of our evaluation, both experts and amateurs found unexpected uses of *SpeechMirror* (VDAR). We provide more details in our supplemental document (Section 5.5) and video.

6.5.1 Expert user: Coaching with SpeechMirror

EE1, a district champion for the Toastmasters speech evaluation contest and a university and youth speech coach noticed from the Factor Panel that the speaker he was analyzing had exceptionally low arousal, most similar to speakers at the club level. He then used the Mirror Panel to find one of the most different SpeechTwins. He noted the speaker was the previous year's world 2nd place winner. He then switched the video to the comparison video and used the timeline to find a moment of especially high arousal. After noticing the changes in arousal over time, he said "*I think the tool could be very useful. In order to build a connection to the audience, you need contrast, you need the twists and turns of emotions. Then you can elicit emotions of the audience. I can use this to show my student that there is strong contrast [for this expert speaker]. I could let the student see where they are and compare with*

different speakers." While EE1's mouse was hovering on a SpeechTwin with a smooth spread of positional footprints, EE1 added, "*With the SpeechTwin we can see where they are and compare with the other speakers to see possibilities.*"

He then moved his mouse to indicate a SpeechTwin with a smooth distribution of footprints and commented "*this speaker had constant stage movement.*" He then pointed at a SpeechTwin with a clear three part distribution of footprints, stating "*if like this there is a clear stage design*". He noted the possible use of our system for coaching, "*Overall I feel like this is a tool that I could use and discuss with my students and find out different insights so that they can improve.*"

6.5.2 Amateur user: Improving with SpeechMirror

EA2 was a previous district contestant that had attended 2-5 online contests. During evaluation, she used our system to better observe changes in her valence and to find more possibilities of improving her gestures. She first noted that while overall, the average valence in her face was indicated by the system as performing well, the system found her speech most similar to someone that was smiling throughout their speech. She guessed it might be that her sadness wasn't expressed obviously enough. When looking at the Time Slice panel, she observed that the ending of the speech was happy, "*This is correct.*" However, she then pointed to earlier time slices in her speech "*For these two parts I really wanted to show my sadness, my anger, but I guess I failed.*"

While comparing by selecting both the valence average and gesture energy change, she stated "*I want to see who is the most similar to me.*" She selected the most similar speech, noting that the gestures, similar to hers, were mostly below the screen or on the screen border.

She then looked at the Mirror Panel to find the most different speaker. When she played the video, she saw a speaker that had many different gestures and stage changes. While looking at the speech comparison board to compare to her speech across factors, she noted "*Ahhh... much more pose diversity than me.*" With the Mirror Panel EA2 was quickly able to find speakers that use gestures very differently, more often over the border of the screen, and with a wider variety.

7 Discussion

In this section, we discuss the lessons learned in our research, the limitations of our work, and the implications for future work.

Analysis of Online Public Speaking. Our system has received high appreciation from domain users for its ability to evaluate and analyze online speech techniques. This demonstrates the value and potential of our system. Users have also expressed their expectation for additional features and capabilities in our system. Most of the participants mentioned that a report of the analysis results can offer an intuitive understanding of a speech, alleviate the analysis burden for users, and provide clear guidance for future exploration. Our system can be further simplified for better user experience by putting less on the interface. Improving text visualization and analysis is a potential area for future work, addressing *SpeechMirror*'s limitations in readability and supporting advanced analysis. This can be achieved through exploring text feature analysis and incorporating text visualizations [28].

Visualization of Human Body Language. We have employed novel approaches to visualize data in different modalities of presentation techniques. Combining multiple modalities of data to provide a unified visualization analysis while preserving the distinctive characteristics of each modality poses significant challenges. Balancing the amount of information and complexity in multimodal visualization requires further consideration. Additionally, the color-emotion mapping scheme in our system may hinder visual accessibility, particularly for individuals with red-green color blindness. To address this, an optional color scheme is provided in Section 4.3 of the supplementary material, enabling individuals with color blindness to utilize our system.

Scope of Presentation Techniques. We aim at providing a comprehensive list of speech techniques for users to explore and improve, which has received positive feedback from users. However, domain experts thought of more techniques that should be taken into consideration. In our evaluation, several expert participants mentioned that background scenery, lighting, and visual aids are also important for online speeches. However, the extraction of these features requires further exploration. We tried to detect prop usage through human-object interaction models [17, 68], but received suboptimal outcomes due to the limited recognizable objects and interaction classes. As the development of relevant models continues, we expect more presentation techniques to be quantified and analyzed in future work.

Recommendation of Speeches. The recommendation function in our system simplifies the process of discovering new videos for users, and has been well-received by our participants. More recommendation approaches can be further considered, such as based on the themes or topics of speeches, on the placements of speeches, or on other tags for users to filter. Moreover, recommendations made by fusing multimodal raw features of presentation techniques could be promising.

Model Accuracy. We utilized state-of-the-art models for multimodal feature extraction. However, apart from the aforementioned limitations regarding the extraction of presentation techniques, we observed inaccuracies of the models in certain situations. For instance, when a webcam is positioned too low, capturing an upward-facing view of the speaker's face, it can result in inaccurate eye gaze and facial emotion recognition. Overall, our models meet practical accuracy requirements, but the performance can be enhanced by incorporating more precise feature extraction models. For the factor effectiveness estimation model, we may explore more precise models or models that consider multiple factors simultaneously. To address potential inaccuracies, we suggest introducing uncertainty visualization in future research. Accurately assessing and visually encoding uncertainty may pose significant challenges.

Volume of Data. *SpeechMirror* is the first visual analytics system that analyzes the effectiveness of speeches and recommends speech samples based on a collection of videos. We anticipate that as the volume of speech video data increases, our effectiveness estimation model will generate more accurate and robust results. With larger volume of video, the system will also provide more diverse recommendation results in both expression and speech content.

Generalization of Our Methods. Our work focuses on online speech contests in the WCPS. Our work can also be extended to other speaking and presentation scenarios in the future. Furthermore, one participant in the evaluation (EE1) suggested that our system could be adapted and applied to other fields that require the analysis of body language, such as behavior analysis during suspect interrogations.

Privacy Issue. *SpeechMirror* provides users with other speakers' speech videos as a basis for analysis, which inevitably raises privacy concerns for the speakers. Therefore, we obtain the consent of the speakers in actual usage and restrict users' access to the data and system. However, when deploying similar systems in public use, ethical considerations should be addressed. It is also important to consider the potential for imitation and plagiarism of presentation techniques with such systems, which requires further exploration in future work.

8 Conclusion

In this paper, we propose *SpeechMirror*, a visual analytics system for public speaking experts and amateurs to evaluate a speech and explore potential improvements. Our system is the first to evaluate speaker performance on a single speech for speaking techniques and provide an in-depth analysis using a comprehensive set of factors, allowing users to identify and understand areas for improvement of their speech. Additionally, the system recommends users speeches by similarity and difference of techniques and content to enhance their understanding of different ways of expression. Novel visualizations are employed for intuitive understanding of presentation techniques. We designed the system based on insights from literature review and domain interviews. A user study was conducted to evaluate the system, indicating the effectiveness of our system in user experience and gaining insights.

In future work, we plan to further improve our system in usability and functionality based on the aforementioned reflections on our system, including (1) making the interface easier to use, (2) expanding the scope of presentation techniques, (3) improving the accuracy of models. Furthermore, we intend to promote our system for applications and collect more speech data for more potential research.

Acknowledgments

We wish to thank our domain collaborators and the anonymous reviewers for their constructive comments. This work was supported by the National Key R&D Program of China (2022ZD0117900) and Beijing Natural Science Foundation (4212029).

References

[1] O. Alemi, P. Pasquier, and C. Shaw. Mova: Interactive movement analytics platform. In *Proceedings of the 2014 International Workshop on Movement and Computing*, MOCO '14, p. 37–42. Association for Computing Machinery, New York, NY, USA, 2014. doi: 10.1145/2617995.2618002 3

[2] O. Arriaga, M. Valdenegro-Toro, and P. Plöger. Real-time convolutional neural networks for emotion and gender classification. *arXiv preprint arXiv:1710.07557*, 2017. 5

[3] T. Baltrusaitis, A. Zadeh, Y. C. Lim, and L.-P. Morency. Openface 2.0: Facial behavior analysis toolkit. In *2018 13th IEEE International Conference on Automatic Face Gesture Recognition (FG 2018)*, pp. 59–66, 2018. doi: 10.1109/FG.2018.00019 5

[4] G. E. Batista, E. J. Keogh, O. M. Tataw, and V. de Souza. Cid: an efficient complexity-invariant distance for time series. *Data Mining and Knowledge Discovery*, 28(3):634–669, 2014. 5

[5] S. Bird, E. Klein, and E. Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. "O'Reilly Media, Inc.", 2009. 5

[6] A. Blair-Early and M. Zender. User interface design principles for interaction design. *Design Issues*, 24(3):85–107, 2008. 5

[7] P. Boersma and D. Weenink. Praat: doing phonetics by computer [Computer program]. Version 6.1.38, retrieved 2 January 2021 http://www.praat.org/, 2021. 5

[8] M. Bostock, V. Ogievetsky, and J. Heer. D³ data-driven documents. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2301–2309, 2011. doi: 10.1109/TVCG.2011.185 4

[9] M. Böttinger, H.-N. Kostis, M. Velez-Rojas, P. Rheingans, and A. Ynnerman. Reflections on visualization for broad audiences. *Foundations of data visualization*, pp. 297–305, 2020. 5

[10] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar, et al. Universal sentence encoder. *arXiv preprint arXiv:1803.11175*, 2018. 5

[11] H. Chernoff. The use of faces to represent points in k-dimensional space graphically. *Journal of the American statistical Association*, 68(342):361–368, 1973. 6

[12] M. Contributors. Openmmlab pose estimation toolbox and benchmark. https://github.com/open-mmlab/mmpose, 2022. 5

[13] S. D'Angelo and D. Gergle. An eye for design: gaze visualizations for remote collaborative work. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–12, 2018. 3

[14] J. Donovan and R. Avery. *Speaker, Leader, Champion: Succeed at Work Through the Power of Public Speaking, featuring the prize-winning speeches of Toastmasters World Champions*. McGraw Hill Professional, 2014. 3

[15] V. Echeverría, A. Avendaño, K. Chiluiza, A. Vásquez, and X. Ochoa. Presentation skills estimation based on video and kinect data analysis. In *Proceedings of the 2014 ACM Workshop on Multimodal Learning Analytics Workshop and Grand Challenge*, MLA '14, p. 53–60. Association for Computing Machinery, New York, NY, USA, 2014. doi: 10.1145/2666633.2666641 1, 2

[16] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, p. 226–231. AAAI Press, 1996. 5

[17] D. C. Frederic Z. Zhang and S. Gould. Spatially conditioned graphs for detecting human–object interactions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13319–13327, October 2021. 9

[18] A. Geitgey. face_recognition. https://github.com/ageitgey/face_recognition, 2018. 5

[19] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, et al. Challenges in representation learning: A report on three machine learning contests. In *International conference on neural information processing*, pp. 117–124. Springer, 2013. 5

[20] E. P. Green. Presenting virtually. In *Healthy Presentations*, pp. 87–100. Springer, 2021. 1

[21] P. A. Gutiérrez, M. Perez-Ortiz, J. Sanchez-Monedero, F. Fernandez-Navarro, and C. Hervas-Martinez. Ordinal regression methods: survey and experimental study. *IEEE Transactions on Knowledge and Data Engineering*, 28(1):127–146, 2015. 5

[22] K. Higuchi, S. Matsuda, R. Kamikubo, T. Enomoto, Y. Sugano, J. Yamamoto, and Y. Sato. Visualizing gaze direction to support video coding of social attention for children with autism spectrum disorder. In *23rd International Conference on Intelligent User Interfaces*, pp. 571–582, 2018. 3

[23] C. Ionescu, D. Papava, V. Olaru, and C. Sminchisescu. Human3.6m: Large scale datasets and predictive methods for 3d human sensing in natural environments. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(7):1325–1339, jul 2014. 5

[24] Y. Jadoul, B. Thompson, and B. de Boer. Introducing Parselmouth: A Python interface to Praat. *Journal of Phonetics*, 71:1–15, 2018. doi: 10.1016/j.wocn.2018.07.001 5

[25] W. Jiang, S. Jin, W. Liu, C. Qian, P. Luo, and S. Liu. Posetrans: A simple yet effective pose transformation augmentation for human pose estimation. *arXiv preprint arXiv:2208.07755*, 2022. 6

[26] K. Kirchhoff and S. Chen. Cmu pronouncing dictionary (cmudict). https://github.com/cmusphinx/cmudict, 2018. Accessed on: January 11, 2021. 5

[27] D. Kravvaris and K. L. Kermanidis. Speakers' language characteristics analysis of online educational videos. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, pp. 60–69. Springer, 2014. 1, 2

[28] K. Kucher and A. Kerren. Text visualization techniques: Taxonomy, visual survey, and community insights. In *2015 IEEE Pacific Visualization Symposium (PacificVis)*, pp. 117–121, 2015. doi: 10.1109/PACIFICVIS.2015.7156366 9

[29] H. Lam, E. Bertini, P. Isenberg, C. Plaisant, and S. Carpendale. Empirical studies in information visualization: Seven scenarios. *IEEE transactions on visualization and computer graphics*, 18(9):1520–1536, 2011. 7

[30] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014. 5

[31] G. Luzardo, B. Guamán, K. Chiluiza, J. Castells, and X. Ochoa. Estimation of presentations skills based on slides and audio features. In *Proceedings of the 2014 ACM Workshop on Multimodal Learning Analytics Workshop and Grand Challenge*, MLA '14, p. 37–44. Association for Computing Machinery, New York, NY, USA, 2014. doi: 10.1145/2666633.2666639 1, 2

[32] K. Maher, Z. Huang, J. Song, X. Deng, Y.-K. Lai, C. Ma, H. Wang, Y.-J. Liu, and H. Wang. E-ffective: A visual analytic system for exploring the emotion and effectiveness of inspirational speeches. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):508–517, Jan 2022. doi: 10.1109/TVCG.2021.3114789 1, 2, 3

[33] S. McLean. *Business Communication for Success*. Business Communication for Success. Flat World Knowledge, 2015. 3

[34] Microsoft. Azure cognitive speech to text service. https://azure.microsoft.com/en-us/services/cognitive-services/speech-to-text/. Accessed March 4, 2022. 5

[35] A. Mollahosseini, B. Hasani, and M. H. Mahoor. Affectnet: A database for facial expression, valence, and arousal computing in the wild. *IEEE Trans. Affect. Comput.*, 10(1):18–31, Jan. 2019. doi: 10.1109/TAFFC.2017.2740923 5

[36] R. Niewiadomski, M. Mancini, and S. Piana. Human and virtual agent expressive gesture quality analysis and synthesis. *Coverbal Synchrony in Human-Machine Interaction*, pp. 269–292, 2013. 5

[37] B. Nojavanasghari, D. Gopinath, J. Koushik, T. Baltrušaitis, and L.-P. Morency. Deep multimodal fusion for persuasiveness prediction. In *Proceedings of the 18th ACM International Conference on Multimodal Interaction*, ICMI '16, p. 284–288. Association for Computing Machinery, New York, NY, USA, 2016. doi: 10.1145/2993148.2993176 1, 2

[38] X. Ochoa, F. Domínguez, B. Guamán, R. Maya, G. Falcones, and J. Castells. The rap system: Automatic feedback of oral presentation skills using multimodal analysis and low-cost sensors. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*, LAK '18, p. 360–364. Association for Computing Machinery, New York, NY, USA, 2018. doi: 10.1145/3170358.3170406 3

[39] O. Pakov and D. Miniotas. Visualization of eye gaze data using heat maps.

Elektronika Ir Elektrotechnika, 115(2), 2007. 3

[40] Pallets. Flask's documentation (1.1.x). https://flask.palletsprojects.com/en/1.1.x/. Accessed Januray 12, 2022. 4

[41] S. Plagenhoef, F. G. Evans, and T. Abdelnour. Anatomical data for analyzing human motion. *Research quarterly for exercise and sport*, 54(2):169–178, 1983. 5

[42] R. L. Potter. 9.5 presenting virtually. *Technical Writing Essentials*, 2022. 1

[43] J. Quoidbach, J. Gruber, M. Mikolajczak, A. Kogan, I. Kotsou, and M. I. Norton. Emodiversity and the emotional ecosystem. *Journal of experimental psychology: General*, 143(6):2057, 2014. 5

[44] V. Ramanarayanan, C. W. Leong, L. Chen, G. Feng, and D. Suendermann-Oeft. Evaluating speech, face, emotion and body movement time-series features for automated multimodal presentation scoring. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, p. 23–30. Association for Computing Machinery, New York, NY, USA, 2015. doi: 10.1145/2818346.2820765 1, 2

[45] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 39(06):1137–1149, jun 2017. doi: 10.1109/TPAMI.2016.2577031 5

[46] M. Rimol. Gartner says the majority of technology products and services will be built by professionals outside of it by 2024, Jun 2021. 2

[47] P. Sanchez. *Presenting Virtually: Communicate and Connect with Online Audiences*. Duarte Guide. DUARTE Press, 2021. 1, 4

[48] R. S. Schaefer, L. J. Beijer, W. Seuskens, T. C. Rietveld, and M. Sadakata. Intuitive visualizations of pitch and loudness in speech. *Psychonomic bulletin & review*, 23:548–555, 2016. 7

[49] J. Schneider, D. Börner, P. van Rosmalen, and M. Specht. Presentation trainer: what experts and computers can tell about your nonverbal communication. *Journal of Computer Assisted Learning*, 33(2):164–177, 2017. doi: 10.1111/jcal.12175 4

[50] R. Sharma, T. Guha, and G. Sharma. Multichannel attention network for analyzing visual behavior in public speaking. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 476–484, 2018. doi: 10.1109/WACV.2018.00058 1, 2

[51] B. Sievers, C. Lee, W. Haslett, and T. Wheatley. A multi-sensory code for emotional arousal. *Proceedings of the Royal Society B*, 286(1906):20190513, 2019. 6

[52] L. South, M. Schwab, N. Beauchamp, L. Wang, J. Wihbey, and M. A. Borkin. Debatevis: Visualizing political debates for non-expert users. In *2020 IEEE Visualization Conference (VIS)*, pp. 241–245, 2020. doi: 10.1109/VIS47514.2020.00055 2

[53] M. Stein, H. Janetzko, A. Lamprecht, T. Breitkreutz, P. Zimmermann, B. Goldlücke, T. Schreck, G. Andrienko, M. Grossniklaus, and D. A. Keim. Bring it to the pitch: Combining video and movement data to enhance team sport analysis. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):13–22, 2018. doi: 10.1109/TVCG.2017.2745181 3

[54] K. Sun, B. Xiao, D. Liu, and J. Wang. Deep high-resolution representation learning for human pose estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5693–5703, 2019. 5

[55] M. I. Tanveer, J. Liu, and M. E. Hoque. Unsupervised extraction of human-interpretable nonverbal behavioral cues in a public speaking scenario. In *Proceedings of the 23rd ACM International Conference on Multimedia*, MM '15, p. 863–866. Association for Computing Machinery, New York, NY, USA, 2015. doi: 10.1145/2733373.2806350 1, 2

[56] M. I. Tanveer, S. Samrose, R. A. Baten, and M. E. Hoque. *Awe the Audience: How the Narrative Trajectories Affect Audience Perception in Public Speaking*, p. 1–12. Association for Computing Machinery, New York, NY, USA, 2018. 1, 2

[57] H.-N. Teodorescu and M. Hagan. Experimental, ad hoc, online, inter-university student e-contest during the pandemic – lessons learned. In *2020 12th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, pp. 1–6, 2020. doi: 10.1109/ECAI50035.2020.9223243 3

[58] T. Tsai. Are you ted talk material? comparing prosody in professors and ted speakers. In *Sixteenth Annual Conference of the International Speech Communication Association*, 2015. 1, 2

[59] X. Wang, Y. Ming, T. Wu, H. Zeng, Y. Wang, and H. Qu. Dehumor: Visual analytics for decomposing humor. *IEEE Transactions on Visualization and Computer Graphics*, pp. 1–1, 2021. doi: 10.1109/TVCG.2021.3097709 2, 3

[60] X. Wang, H. Zeng, Y. Wang, A. Wu, Z. Sun, X. Ma, and H. Qu. *VoiceCoach: Interactive Evidence-Based Training for Voice Modulation Skills in Public Speaking*, p. 1–12. Association for Computing Machinery, New York, NY, USA, 2020. 2, 3

[61] E. Wood, T. Baltruaitis, X. Zhang, Y. Sugano, P. Robinson, and A. Bulling. Rendering of eyes for eye-shape registration and gaze estimation. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pp. 3756–3764, 2015. doi: 10.1109/ICCV.2015.428 5

[62] T. Wörtwein, M. Chollet, B. Schauerte, L.-P. Morency, R. Stiefelhagen, and S. Scherer. Multimodal public speaking performance assessment. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, p. 43–50. Association for Computing Machinery, New York, NY, USA, 2015. doi: 10.1145/2818346.2820762 1, 2

[63] A. Wu and H. Qu. Multimodal analysis of video collections: Visual exploration of presentation techniques in ted talks. *IEEE Transactions on Visualization and Computer Graphics*, 26(7):2429–2442, July 2020. doi: 10.1109/TVCG.2018.2889081 1, 2, 3

[64] L. Yuan, Y. Chen, S. Fu, A. Wu, and H. Qu. Speechlens: A visual analytics approach for exploring speech strategies with textural and acoustic features. In *2019 IEEE International Conference on Big Data and Smart Computing (BigComp)*, pp. 1–8, Feb 2019. doi: 10.1109/BIGCOMP.2019.8679261 2

[65] H. Zeng, X. Shu, Y. Wang, Y. Wang, L. Zhang, T.-C. Pong, and H. Qu. Emotioncues: Emotion-oriented visual summarization of classroom videos. *IEEE Transactions on Visualization and Computer Graphics*, 27(7):3168–3181, 2021. doi: 10.1109/TVCG.2019.2963659 3

[66] H. Zeng, X. Wang, Y. Wang, A. Wu, T.-C. Pong, and H. Qu. Gesturelens: Visual analysis of gestures in presentation videos. *IEEE Transactions on Visualization and Computer Graphics*, 2022. 2, 3

[67] H. Zeng, X. Wang, A. Wu, Y. Wang, Q. Li, A. Endert, and H. Qu. Emoco: Visual analysis of emotion coherence in presentation videos. *IEEE Transactions on Visualization and Computer Graphics*, 26(1):927–937, Jan 2020. doi: 10.1109/TVCG.2019.2934656 2, 3

[68] F. Z. Zhang, D. Campbell, and S. Gould. Efficient two-stage detection of human-object interactions with a novel unary-pairwise transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20104–20112, June 2022. 9

[69] J. R. Zhang. Upper body gestures in lecture videos: Indexing and correlating to pedagogical significance. In *Proceedings of the 20th ACM International Conference on Multimedia*, MM '12, p. 1389–1392. Association for Computing Machinery, New York, NY, USA, 2012. doi: 10.1145/2393347.2396499 1, 2