DeHumor: Visual Analytics for Decomposing Humor

Xingbo Wang, Yao Ming, Tongshuang Wu, Haipeng Zeng, Yong Wang, and Huamin Qu

Abstract—Despite being a critical communication skill, grasping humor is challenging—a successful use of humor requires a mixture of both engaging content build-up and an appropriate vocal delivery (e.g., pause). Prior studies on computational humor emphasize the textual and audio features immediately next to the punchline, yet overlooking longer-term context setup. Moreover, the theories are usually too abstract for understanding each concrete humor snippet. To fill in the gap, we develop *DeHumor*, a visual analytical system for analyzing humorous behaviors in public speaking. To intuitively reveal the building blocks of each concrete example, *DeHumor* decomposes each humorous video into multimodal features and provides inline annotations of them on the video script. In particular, to better capture the build-ups, we introduce content repetition as a complement to features introduced in theories of computational humor and visualize them in a context linking graph. To help users locate the punchlines that have the desired features to learn, we summarize the content (with keywords) and humor feature statistics on an augmented time matrix. With case studies on stand-up comedy shows and TED talks, we show that *DeHumor* is able to highlight various building blocks of humor examples. In addition, expert interviews with communication coaches and humor researchers demonstrate the effectiveness of *DeHumor* for multimodal humor analysis of speech content and vocal delivery.

Index Terms—Humor, Context, Multimodal Features, Visualization.

✦

1 INTRODUCTION

HUMOR—the use of puns, turns of phrases, or humorous anecdotes—is a powerful communication skill for public speakers to connect, engage, and entertain their audiences. A proper usage of humor can help induce shared amusement [1], reduce social anxiety [2], and boost persuasive power [3]. Although *identifying* humor (i.e., judging whether a joke is funny or not) comes natural to us, *becoming* humorous is challenging in practice, as it requires the integration of various humor skills. To create humorous content (i.e., jokes), the speakers need to come up with intelligent gradual setups, as well as a sudden twist to subvert the audience's expectation [4]. Then, to effectively achieve the intended dramatic effect, speakers have to decorate the contents with appropriate acoustic delivery methods [5], [6] (e.g., pause, pitch). Thorough understanding and learning of humor can only be achieved if we can decompose these building blocks and the interactions between them.

Prior work across multiple disciplines (psychology, philosophy, and linguistics) has qualitatively characterized humor. For example, Plato described humor as an expression of superiority to others, while Schopenhauer [7] stated that humor comes from the realization of incongruous interpretations of a statement. These abstract theories become the cornerstone of various computational features that capture phonetic [8], [9], stylistic [8], [10], human-centered [9], [11], [12], and content-based [13], [14] humor characteristics.

While these features make it possible to quantify humor, analyzing concrete humor examples in speeches remains challenging for two reasons. First, presenting a laundry list of features for a humorous speech can be overwhelming. Not only do the features create a heavy perceptual burden, but the sophisticated interactions between different building blocks remain hidden in the large feature space. Most of the analyses to date [8], [9], [11], [15], [16], [17], [18] only focus on either humorous content or vocal delivery independently. However, both of them are needed simultaneously to understand humor in practice. A **punchline**—the most important sentence that triggers the audience response (e.g., laughter)—may be mundane in isolation but hilarious when rendered with exaggerated volume and pitch. In the example[1] below, patterns like acoustic stressing at the modal particles (i.e., getting louder at *"okay"* in Line #5) or pausing to emphasize turning points (i.e., pausing before *"I"* in Line #4 to contrast with *"they"* in the preceding sentence) are only observable when we highlight their occurrences.

1	*So when I show up to a crime scene,*
2	*Somebody is always like, "are you a cop?"*
3	*I don't wanna say I'm a cop cause it's against the law.*
4	*So they go, "are you a cop?"*
5	*And I go, [PAUSE] "I'll ask the f**king questions, okay [LOUDER]?"*

Second, the already overwhelming feature definitions have yet to be comprehensive. Existing research emphasizes short jokes (e.g., one-liners) while overlooking those with longer-term set-ups, making it difficult to track the clues that help lead to the core punchline. In the previous example,

- X. Wang and H. Qu are with the Hong Kong University of Science and Technology.
  E-mail: {xwangeg, huamin}@cse.ust.hk
- Y. Ming is with Bloomberg LP.
  E-mail: yming7@bloomberg.net
- T. Wu is with the University of Washington.
  E-mail: wtshuang@cs.washington.edu
- H. Zeng is with the Sun Yat-sen University.
  E-mail: zenghp5@mail.sysu.edu.cn
- Y. Wang is with the Singapore Management University.
  E-mail: yongwang@smu.edu.sg

1. The example link: http://bit.ly/2MrVKf9

the punchline in Line #5 only becomes funny after Lines #1 to #5 provide the essential context: When attempting to enter a crime scene (#1), the speaker was asked about the cop identity (#2). Not being an actual cop, he avoided explicitly making an illegal claim that he is one (#3). As a workaround, he quoted a common trope for police in movies and TV (*I'm asking the f**king question!*, #5), and so to mislead the *"somebody"* to believe that he is a cop. This example demonstrates the importance of contexts for humor analysis, which motivates our study.

To better decompose humor examples into critical features, as well as intuitively present the mixtures of these features, we present a novel visual analytics system named *DeHumor*. *DeHumor* aims to help domain experts (e.g., communication coaches and researchers) analyze the verbal content and vocal delivery of public speaking containing many humorous punchlines (labeled with audience response markers like `[LAUGHTER]`). We formulate the design requirements based on literature surveys and user interviews with five humor researchers as well as two communication coaches. We choose to interview these experts because they have theoretical knowledge in humor and need assistance on a systematic investigation of humor. Accordingly, we design *DeHumor* to support multi-level explorations of humorous texts and delivery in speeches. We aim to enable users to easily understand when and where humorous punchlines are inserted (speech-level), how one particular punchline relates to its preceding build-up sentences (context-level), and how the vocal delivery and the textual content are paired within each sentence (sentence-level). In particular, to reveal the interactions between textual and audio features, we provide inline annotations of the features along with the raw transcripts. To highlight the build-ups, we introduce a context linking graph that can recognize relevant phrases and visually connect them with links. With case studies on stand-up comedy shows and TED talks, we show that *DeHumor* can highlight various building blocks of humor examples. Interviews with domain experts further confirm that *DeHumor* is helpful for exploratory and in-depth analysis of humorous snippets.

In summary, the major contributions of our work are:

- We design a visual analytics system to support interactive and multimodal analysis of humorous pieces and reveal humor strategies in speech content and voice.
- We demonstrate the usability and effectiveness of *DeHumor* through case studies and expert interviews with communication coaches and humor researchers.

2 RELATED WORK

This section reviews related research on humor theory and visualization for speech analysis.

2.1 Computational Humor Features on Speech

Modeling humor features is beneficial and critical for automatic humor understanding. Prior work has modeled humor using both text and audio features. For textual features, Mihalcea and Strapparava [8] extracted stylistic features that characterize humorous texts, including alliteration, antonym, and adult slang. Later, Mihalcea and

Pulman [11] extended feature sets with human centeredness and polarity orientation. Kiddon and Brun [19] measured erotica-level of nouns, adjectives, and verb phrases. Zhang and Liu [12] designed five categories of humor-related linguistic features, including morpho-syntactic features, lexico-semantic features, pragmatic features, and phonetic features [9], [20]. Content-based features (e.g., n-grams [21], [22], lexical centrality [14], incongruity [9], and word associations [23]) were also widely experimented to study the patterns in humorous text content in previous work. However, most of these features were not systematically derived and were defined in an empirical way. Yang et al. [9] proposed a computational framework to describe the latent semantic structures of humor, including incongruity structure, ambiguity theory, interpersonal effect, and phonetic style. Bali et al. [24] extracted three major characteristics across all humor types, which are mode, theme, and topic. The mode (e.g., exaggeration) is dependent on situations of delivery. The theme relates to emotions behind the use of language. The topic covers the central elements of humor. In our work, we integrate and extend the above frameworks for textual features to analyze humorous texts.

For audio features, previous quantitative studies [15], [16], [17], [18], [25], [26] identified significant dimensions for joke-telling, including volume, pitch, speech rate, and pause length. Pickering et al. [15] found punchlines were produced with lower pitch in joke narrative. Attardo and Pickering [17] investigated pauses around punchlines. Purandare and Litman [25] used acoustic-prosodic features (i.e., pitch, energy, and tempo) and linguistic features to automatically recognize the humor in the TV sitcom. Bertero and Fung [26] modeled conversational humor by combining speech utterances with a set of high level features (e.g., speaking rate). Our work computes pitch, volume, speed, and pauses around punchlines and their contexts, to reveal acoustic patterns in the humor delivery.

2.2 Visualization for Speech Analysis

Speech visualization is an important research topic in the multimedia analysis. It is applied in many domains, such as public speaking training [27], [28], visualization for the hearing impaired [29], and emotion analysis [30]. While some prior studies have visualized the speaker/audience interactions and topic dynamics in multi-party speeches (e.g., debate, conversation) [31], [32], [33], our work focuses on analyzing verbal content (e.g., word use) and vocal delivery (e.g., voice modulation) of humor in public speaking.

One of the main goals of visualizing speech data is to intuitively and effectively reveal the relationship between content and speaking voice. The most straightforward way is to encode sequential features as bar charts or line charts and then draw them along the script [30], [34]. Oh [35] used a vertical timeline to summarize features of sections in songs. However, directly overlaying features on the words can lead to a high cognitive load. Moreover, it does not explicitly demonstrate relationships between words. Patel and Furr [36] proposed a method to directly encode the prosodic features using text properties. It manipulates the vertical offset, opacity, and letter spacing of texts to represent pitch, intensity and audio duration, respectively. Similarly, Wang

et al. [27] and Rubin et al. [28] designed intuitive glyphs to represent prosodic features, which annotates speakers' vocal performance on the script. Similarly, in our work, we design glyphs and adjust text styles to explicitly demonstrate the humor features and their relationships.

Besides, we aim to visualize the semantic relationship of texts in humor snippets to help understand textual humor. Here, we summarize the prior studies that have inspired our research. Matrices [37], [38] are widely adopted to visualize co-occurrence patterns in text documents. Word clouds [39] can also summarize word relations. It is also common to use graphs [40], [41] and links [42], [43], [44] to describe co-occurrence and repetitions. However, it is challenging to directly apply these techniques in our work. For example, word clouds lose temporal information. Matrices suffer from space inefficiencies, especially when they are sparse. Arc diagrams alleviate the above issues by placing words in a line and visually connecting them. Still, if the text is long, it is difficult to obtain an overview of the text relationship while keeping the temporal orders. In this paper, we extend the arc diagram with a multi-level context summary and rich interactions to support the effective identification of contextual repetitions in a humor snippet.

3 DESIGN PROCESS

Our goal is to support an in-depth and systematic investigation into the humor composition and its vocal delivery in public speaking (e.g., oral presentation). Our main target users are people who have theoretical knowledge in humor and are motivated to study humorous speech (e.g., humor researchers and communication coaches). We expect our system to alleviate their mental burden when analyzing unstructured humorous speech (i.e., texts and audio) in an organized and quantitative way.

To build a concrete understanding of humor, we conducted literature reviews and informal user interviews to identify a set of textual and audio features that are both quantifiable and essential for humor analysis. Specifically, we first summarized features from existing literature and proposed a new computational method for extracting the context-related feature (i.e., inter-sentence repetition) to supplement the framework of computational humor.

Next, based on these feature candidates, we interviewed five humor researchers and two communication coaches who provided professional insights into humor study and helped validate our proposed features regarding their importance and helpfulness for humor analysis. Meanwhile, during interviews, we inquired about their perspectives on the analytical aspects and challenges within humor analysis. Based on their feedback, we distilled design requirements, which guided our initial system design. The researchers (*E1-E5*, including three postgraduates, one PhD graduate, and one university lecturer) study English/applied linguistics, English literature, and L2 learning. Three of them have published research papers on humor. They all have done humor research projects. The communication coaches (*C1, C2*) have five and ten years of communication skills training experience, respectively. One part of their work is to train speakers to deliver humorous speeches based on pre-selected humor examples or topics.

3.1 Literature Review & User Interviews

3.1.1 Humor features

We borrowed the most common and essential quantifiable features of humor content creation and vocal delivery from the existing work mentioned in Sec. 2.1. For the textual features, we organized and selected our features based on the frameworks proposed by Yang et al. [9] and Bali et al. [24], such that our features cover both semantic structures (e.g., incongruity and phonetic style) and content understanding (e.g., topic). Similarly, for the audio features, we collected features and merged the related ones from different studies (e.g., tempo [50] v.s. speech rate [15], and energy [50] v.s. volume [15]). As a result, we covered four audio aspects: volume, pitch, pause, and speed. The full list of features is in Table 1, and the computations are in Sec. 5.1.

While these features comprehensively summarize different aspects of one-liners, existing computational research rarely covers context features in humor cases with longer set-ups. For example, **(inter-sentence) repetition** is one essential comedic devices [51]. Consider a simple example in [52]: *"A: Rover (a dog) is being good. B: I know. C: He is being hungry."* The repetition of the structure "he is being" makes the audience expect a similar response to A's. However, the word "hungry" conflicts with the expectation, and the repetition enhances the dramatic effect of the twist. To seize such patterns in the build-up of a humorous story, we introduce an algorithm to compute context-level *"repetition"*. The detail of the computation is illustrated in Sec. 5.1.3.

3.1.2 User interviews

To validate the proposed humor feature sets and discover analytical needs for humor analysis, we conducted independent interviews with the seven target users (*E1-E5, C1, C2*) mentioned earlier. During interviews, we asked the participants to (1) describe their general process/methodology of humor analysis, (2) illustrate what aspects of humor in speeches they care about and how do they rank our proposed features regarding the importance/difficulty for analysis, (3) propose desired design requirements (e.g., functions) for a system that facilitates systematic humor analysis. Their feedback is reported as follows. The design requirements are summarized in Sec. 3.2.

Whole-to-part analysis. According to the participants' feedback, they generally analyze the speech *from the whole (e.g., speech topics) to the parts (e.g., word use).* They usually first search for humor examples from public speeches, TV shows, and books according to humor topics, genres, and comedians. Then, they focus on the humorous pieces that can elicit laughter from the audience and investigate the patterns of speech content and delivery in humor speeches. Specifically, the analysis follows the **language strata** [53], [54]—*the context, semantics/pragmatics, lexemes (words and phrases), and phones*—from coarse to fine.

Analytical aspects and computational features. As shown in Table 2, *word usage* (rank: 1.86) and *vocal delivery* (rank: 2.57) with the highest importance rankings were considered essential for humor analysis. The participants appraised the extraction of incongruous words, affective lexicons (i.e., sentiment and subjectivity), and phonetic styles (i.e., alliteration and rhyme). These features cover

TABLE 1
A summary of humor-related features that we identify from the qualitative and quantitative research of humor.

	Humor-related features	Subcategory	Description	References
Textual	Content-related features		Key concepts (e.g. situation) on which the humorous story is built	[8], [9], [24], [39]
	Incongruity	Disconnection	Semantic disconnection (e.g., contrast) between two content words in a sentence	[8], [9], [14], [45]
		Intra-sentence repetition	Repeating similar objects in a sentence	[8], [9], [14], [39], [45]
	Human-centeredness	Polarity	Positive/negative orientation of emotion	[8], [9], [12], [45], [46]
		Subjectivity	Subjective/objective orientation	[9], [45], [47], [48]
	Phonetic style	Alliteration	Occurrences of the same letter or sound at the beginning of a group or words	[8], [9], [12], [20], [45]
		Rhyme	Repetition of similar sounds in the final stressed syllables of a group of words	[9], [12], [20], [45]
	Build-ups	Inter-sentence repetition	Concepts (e.g., a person) that have been previously told before a punchline	[4], [6], [49]
Audio	Volume	Volume variation	Variation in volume: softer and louder	[16], [25], [26], [50]
	Pitch	Stress	Vocal stress by raising pitch	[15], [16], [25], [26], [50]
	Pause		A temporary stop in speech	[16], [25], [26], [50]
	Speed	Speed variation	Variation in speech rate: faster and slower	[16], [25], [26], [50]

TABLE 2
The average importance/difficulty rankings for the analytical aspects (A smaller rank value means greater importance/difficulty).

	Word usage	Vocal delivery	Build-ups	Timing	Topics
Importance	1.86	2.57	3.00	3.14	4.43
Difficulty	2.71	2.43	1.86	4.57	3.43

their typical analytical interests and provide quantitative and concrete evidence for language patterns of humor in semantics, lexemes, and phones. *E1* claimed that the incongruous words can reflect the unexpected conflicts and twists in punchlines, which are at the core of an influential humor theory—incongruity. *E3* added that the negative sentiment lexicons help study the styles of self-deprecating or self-enhancing humor. The participants also thought the acoustic features—volume, pitch, pause, speed—can effectively reflect the vocal characteristics of humor. For example, smart pauses (e.g., comic timing) are effective for building up suspense. C1 said that "I can use them (acoustic features) to tell whether a speaker is good at telling jokes or not."

Besides, the two coaches attached much importance to the *timing of humor*. They considered it as a good starting point to learn humor in public speaking. *C1* reasoned that finding a proper place (e.g., speech opening) to insert humor may be the easiest thing for ordinary people to learn, which can make a big impact on their speeches. *E4* suggested the timing should include the distribution and drift of topics ("*What content it supports* and *how the topics evolve*").

In terms of difficulty (Table 2), *humor build-ups* was regarded as the most difficult aspect with the top rank (1.86). The participants thought that the cognitive load of backtracking and comprehending related concepts (e.g., background, characters, emotion) in humor build-ups can be heavy. The inter-sentence repetition extraction was considered reasonable and helpful. *E4* said, "*It (the context repetition) connects the dots (of humor)."* *E3* specified, "*It is useful for revealing the humor structure and can help summarize the core idea of humor."* Still, some context-related humor characteristics proposed by the participants, such as the social background, culture, humor genres (e.g., dark humor), are difficult or unreliable to be quantified. Thus, they are left as future research.

Besides the humor features for word usage, speech content, and vocal delivery (Sec. 2.1), we enrich the existing computational framework with inter-sentence repetitions for humor context analysis and the timing of humor based on the participants' feedback. Their computation and visualization are described in Sec. 5.1.3 and Sec. 5.2 respectively.

Desired functionality. Since there are few tools that enable humor exploration and analysis in various speeches, both coaches and researchers need to manually select and digest speech examples of their interests. It is ineffective and challenging for them to analyze humor in terms of both speech content and vocal delivery. They valued our attempt to build an interactive tool that systematically organizes these computational multimodal features and provides concrete examples to verify the existing humor theories or reveal new insights into humor. We distilled the corresponding design requirements in Sec. 3.2.

3.2 Design Requirements

According to the whole-to-part analysis regarding the language strata, our analytical system should support the hierarchical exploration of humor at different levels—speech level, context level, and sentence level. We summarize the design requirements on these levels as follows.

R1: Analyze text and audio simultaneously to reveal their correlations. Our experts confirmed that both speech content and vocal delivery are considered necessary for a humorous effect. It is difficult to capture both of them by watching the videos. Therefore, at each level, the system needs to present textual and audio features concurrently to help users reason about the effective use of words and voice.

R2: Visualize a speech level overview that shows vocal and verbal styles of humor, as well as their distribution. At the **speech level**, the system should summarize the timing of humor-related properties—the humor is injected how frequently, under what condition (Or, what topic flow), to which part of the speech, and with what verbal and vocal styles. The visual summary of these properties serves as guidance and should help users find specific humor snippets within a speech. For example, a communication coach might prioritize the very first humorous punchline (*when*), to show students how to provide an impressive opening (*objective*) by making small talks or sharing personal lives (e.g., "*My brother's in prison."*). Besides, as suggested by the experts, the visual summary of the humor distribution should be integrated with temporal information, along with the topic flow and verbal feature statistics.

R3: Provide a context-level overview that shows build-up elements of humor, as well as their relationships.

Once zoomed in to a specific snippet, **context level** exploration is necessary for evaluating how a humorous story is written (e.g., how the key concepts in the punchline are first introduced and how they connect the pieces of humor stories), as well as a summary of delivery skills that are frequently used to help convey the story. Both researchers and communication coaches viewed the contextual analysis of humor build-ups to be the most demanding. Therefore, our system should primarily support users at this level.

R4: Highlight the pairing of individual content words and humor-related verbal delivery units. We need to expose the co-occurrence between textual and audio features within each **individual sentence**, so to demonstrate the humor strategies with relevant concrete examples (e.g., words and utterance). Within a snippet, the punchline is the most important sentence since it immediately triggers laughter.

R5: Support intuitive interactions for helping users traverse among different levels, and reveal the different level of details. For example, our communication coaches suggested that the original audio and scripts of the speech should also be included in the system, in addition to a visual summary of humor. The system should allow users to rapidly locate the segments of interest in the speech and playback the corresponding audio clips.

4 SYSTEM OVERVIEW

Motivated by the design requirements, we design and implement an interactive visual analytics system, *DeHumor* (Fig. 1), to explore and analyze verbal humor in public speaking. It combines multimodal humor features with visualization to facilitate users with insights into writing and delivering humor at three levels: speech level, context level, and sentence level.

DeHumor contains three major modules: a data processing module, an analytics module, and a visual interface. The **processing module** extracts humor snippets with aligned audio and transcripts from raw speech videos to support multimodal analysis at different granularities. The **analytics module** computes multimodal humor features from audio and text, which characterize abstract and complex humor behaviors quantitatively. The **interface** visualizes the features extracted by the **analytics module** to support intuitive exploration. Here, we describe our processing module and then provide a brief overview of the interface. We delay the feature extraction in the analytical module, as well as their visual encodings, to Sec. 5.1.

4.1 Data Preprocessing

Data collection. Given a humorous speech, we collect four kinds of data from it: (1) We collect the meta-information (e.g., title, speakers, and categories) for indexing and querying a specific speech, so to enhance the usability of *DeHumor*; (2) We label humor occurrence within a speech based on the audience behavior markers (i.e., [LAUGHTER]) that are annotated in the transcripts; Previous studies have verified that laughter can reliably indicate humor [49], [55], [56], [57]; (3) We use the *transcripts* for content analysis, and (4) the *audio sequences* for verbal delivery analysis.

For demonstration purpose, we collect two speech datasets from *TED Talks* [2] and *Comedy Central Stand-up* [3], which will be described in detail in Sec. 6. Users can prepare speech datasets of similar structures for their interests.

Preprocessing. We process the collected data such that (1) the text script and audio are aligned to support multimodal analysis, and (2) the full speech data is segmented into *humor snippets* to support context and sentence-level analysis. To achieve the alignment, we first detect each word's starting time and ending time in the transcript using P2FA [45]. Thereafter, we align the audio and text modality together at the word level. As for humor snippet segmentation, we regard a sentence immediately before a laughter marker as a **punchline** (i.e., the most important sentence that triggers the audience response). We treat all of the sentences between two punchlines as the **candidate context paragraph** for the second punchline. The intuition is that all the information that occurs after a punchline are potentially useful for building up the next punchline. More concrete context recognition (shown in Sec. 5.1.3) should come from these candidate sentences. As a result, we split the transcripts at laughter markers. Each resulting **humor snippet** contains exactly one punchline (i.e., the last sentence of the segment) and its contexts (all the preceding sentences). The audio is clipped correspondingly through the starting and ending times of the sentences. Eventually, we organize the raw speech data into aligned audio and transcripts per snippet, per sentence, and per word.

4.2 Interface Overview

The user interface follows an overview-to-detail flow. In a collapsible *control panel* (Fig. 1A), a user can use the metadata (name, views, etc.) and the temporal distribution of punchlines (visualized as a **bar code chart** (in Fig. 1A1)) to find their speech-of-interest, which will be loaded in the main component, *humor exploration* view (Fig. 1C). *Humor exploration* visualizes the humor-related details of a speech at different granularity. First, an augmented time matrix (on the left) summarizes the overall patterns of speech topics, humor insertion, and vocal delivery (*R1, R2*). With queries on the time matrix (*R5*) or in the *humor focus* (Fig. 1B), a user can locate a specific humor-snippet-of-interest into the context panel (on the right of Fig. 1C). Within each snippet, the user can examine the humor context (*R3*) through the context linking graph, and understand the interactions between text and audio through the inline humor feature annotations among the transcripts (*R4*). Additional interactions on finding specific context links, related queries, etc. would further support users' exploration experience (*R5*).

5 DEHUMOR

We describe the *humor exploration* view of *DeHumor* in a bottom-up manner (Fig. 1C). First, we illustrate the extraction and encoding of computational humor features in the **sentences and contexts**. Then, the visual summary of the **whole speech**, as well as interactive features of the system, will be explained in detail.

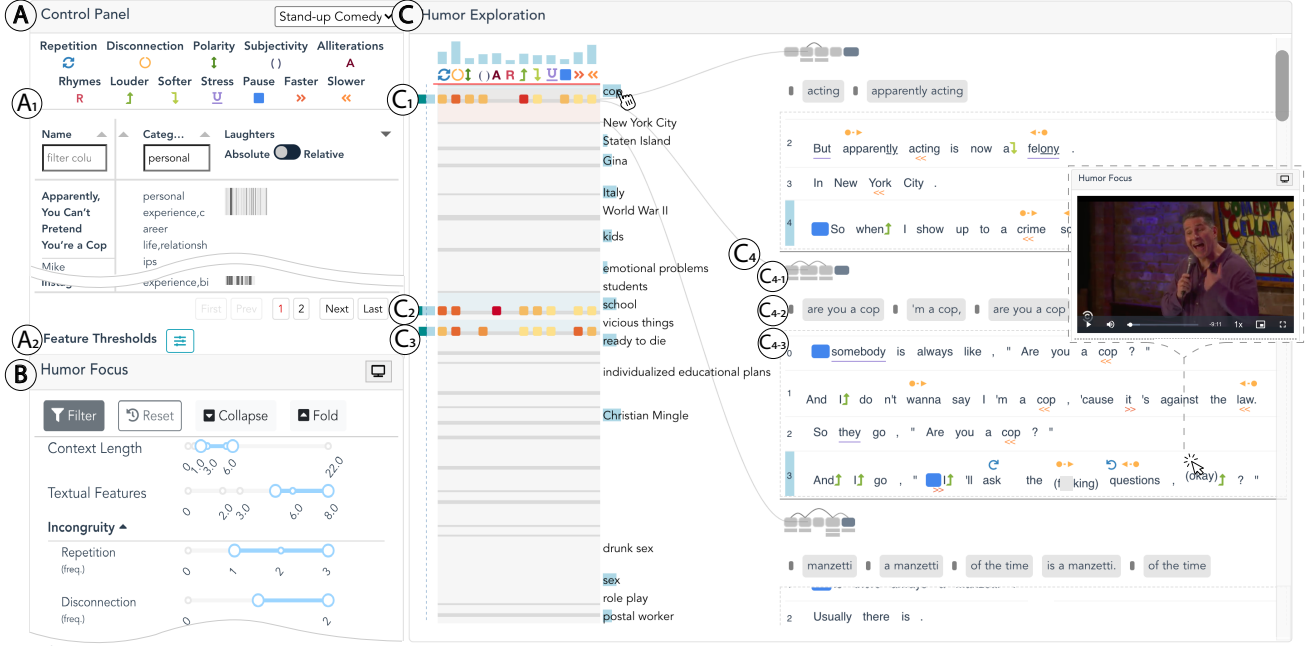2. https://www.ted.com/talks
3. http://www.cc.com/shows/stand-up

Fig. 1. Interactive exploration of speech content and its vocal delivery of humor using DeHumor. The user uses *control panel* (A) to find a speech of interest. The *humor focus* (B) helps the user to further narrow down to the humor snippets with certain verbal and vocal styles. The *humor exploration* (C) guides the multi-level exploration through an augmented time matrix (on the left) for summarizing humor features and context linking graphs (on the right) for analyzing humor context and its speech content and vocal delivery patterns. The user can click on the sentences to show and play the orginal video clips in the *humor focus* (B).

5.1 Humor Feature Analysis and Encoding

We utilize computational humor features to guide and enhance users' reasoning about the styles of verbal humor. First, we describe how the textual (Sec. 5.1.1) and audio (Sec. 5.1.2) features in Tab. 1 are defined, computed, and represented at the **sentence level** (*R2*). To emphasize their potential co-occurrence, we encode all the features with inline glyphs (*R1, R4*). To further facilitate the **context** analysis (*R3*), we design tools for extracting and visualizing the relationship among humor build-ups (Sec. 5.1.3).

5.1.1 Language Features and Glyphs

We compute and encode three types of semantic features at the sentence level (*R4*): incongruity, sentiment, and phonetics. For each feature, a meaningful threshold is used to identify important words or phrases in the sentence, which are annotated with intuitive glyphs. These thresholds can be interactively adjusted by users according to the feature distribution in Fig. 1A2.

Incongruity. Contrasting incongruous concepts (e.g., "clean desk" and "cluttered desk drawer" [8]) is classic for achieving the comic effect. The semantic incongruity of a sentence can be modeled by the repetition and disconnection, or the relative semantic similarities between word pairs [9]. **Disconnection** captures the least semantically similar word pair in the sentence. As shown in Fig. 2A, a pair of dashed arrows (●··▶, ◀··●) are placed above the words "brother" and "prison" to indicate their disconnection. In contrast, **intra-sentence repetition** focuses on the most similar pair. In Fig. 2B, a pair of curved arrows (↻, ↺) show the repetition of the two "cousin"s. The arrows in a pair are pointed to each other, showing the sequential orders and positions of the corresponding word pair in the sentence. We calculate semantic similarity using the cosine

distance on the GloVe [58] embedding. At the sentence-level, we also annotate the sentences that have word pairs with strong disconnections (◯) or repetitions (↻).
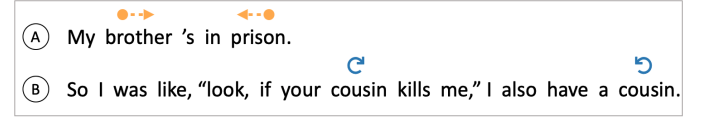

Fig. 2. Examples of incongruity: (A) Disconnection (at beginning of a speech) and (B) Intra-sentence repetition.

Sentiment. Expressing strong sentiment using polarized expressions (how emotionally positive and negative) and subjective statements (how personal) enables a speaker to empathize with the audience. The **polarity** includes both the sentiment direction and sentiment intensity. We use vertical offsets to indicate words with strong polarity. For example, "stupid" in Fig. 3 has a negative polarity, and is therefore displayed with a downside vertical offset. The **subjectivity**, on the other hand, is shown with brackets "( )" around a word. As shown in Fig. 3, "stupid" is associated with the speaker's subjective opinion. We measure word-level and sentence-level polarity and subjectivity using the resource of word annotations and clues for sentiment in [59].
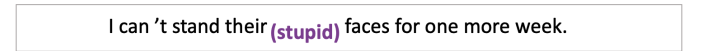

Fig. 3. An example of sentiment expression.

Phonetic style. Phonetic style is often used to achieve catchy verbal deliveries, making the comic effect more memorable and engaging [8]. The most common techniques include (1) **alliteration chains**, which denote multiple words that begin with the same phones, and (2) **rhyme chains**, which include words ending with the same syllables. We utilize the CMU Pronouncing Dictionary [4] within every sen-

4. http://www.speech.cs.cmu.edu/cgi-bin/cmudict

tence, and visually underline the corresponding characters that are responsible for creating the chain.

> Ⓐ God makes <u>dirt</u>, <u>dirt</u> <u>don</u> 't hurt.
>
> Ⓑ I assume your name is <u>Gina</u>, or <u>Regina</u>, or some variation of a saint.

Fig. 4. Examples of phonetic styles. A: Alliteration. B: Rhyme.

5.1.2 Audio Features and Glyphs

We extract and encode the following four representative audio features that reveal the speaker's vocal delivery style. Most of these features are captured by computing its relative significance within a sentence or paragraph. For example, in the speed variation below, we define words to be significantly faster, if it is N times faster than the average of the speed in a sentence, or M times of standard deviations away from the average, with given thresholds N and M (defaulted to 1 and 1.5 respectively).

Speed variation. We compute the Syllables Per Minute (SPM) for each word, as well as the average and standard deviation (SD) of SPM for each sentence. As shown in Fig. 5A, we encode the words which are significantly faster than the sentence as "faster (»)". Similarly, the words which are significantly slower will be labeled as "slower («)".

Pause. We calculate the time intervals between two words. If the interval exceeds a threshold (that defaults to 0.5s), a dark blue rectangle will be drawn in front the corresponding word (e.g., "So" in Fig. 5B). The width of the rectangle encodes the pause length.

Volume variation. We mark the words that are significantly louder or softer than the preceding word in the sentence. They are labeled as "louder (↥)" or "softer (↧)".

Pitch stress. Similar to the volume variation, we derive words that are significantly higher pitched or have more pitch variation based on the pitch contours, and encode them with "stress (U)".

The thresholds above are set according to [27], [28]. They are fine-tuned empirically by testing on audio samples of speech data and can be interactively adjusted in Fig. 1A.

> Ⓐ I 'm ready to die.
>   »»   ««
> Ⓑ ▬▬▬ So when↥I show up to a crime scene.↧
>   ««

Fig. 5. Examples of vocal delivery styles. A: Speed variations. B: A combination of pause, and volume and speed variations.

5.1.3 Humor Context Analysis and Linking

To reveal the relationship among build-ups of a punchline (**R3**), we extract and link similar concepts in the punchline context. As mentioned in Sec. 3.2, a speaker would more frequently repeat useful concepts to help prepare the audience for the upcoming punchline. In the example below, the core takeaway of the punchline is that Germany *does not* have *"fantastic food"*. The message becomes clear because of several repetitions in preceding lines. First, the speaker emphasizes his/her focus on the two countries by repeating (*"the Italian community"*, *"Their people"*, *"Italy"*) and (*"Germany"*) in several places. Second, in Lines #3 and #5, the different modifiers *"a"* and *"no"* before the repeated *"stigma for (being) evil"* highlights the opposite reputations

of Germany and Italy after WWII, and therefore builds a natural comparison between the two. The comparison is then carried on to the punchline, implying the German food is the opposite of Italian's.

> 1 *Let me go after* the Italian community.
> 2 Their people *get off easy.*
> 3 Germany *has a stigma* for being evil.
> 4 *But if you check history,* Italy *fought right alongside Germany in WWII.*
> 5 *But we* have no stigma *for evil, and do you guys know why?*
> 6 *It's because we have fantastic food.*

With this example, we first present an algorithm that captures such inter-sentence repetitions and then describe the visual display.

Concept Grouping Algorithm: Concept grouping may sound trivial at first glance—Naive string match among different tokens may suffice if we assume concepts are always repeated in strictly identical forms. However, in practice, we frequently observe context rephrasing. For example, while the concept entity *"Italy"* is introduced as a modifier for its community in the first sentence, in Line #2 it is just implicitly referred by a pronoun. Beyond entities, more concepts appear in the form of modifier segments (synonym adjectives, similar prepositions on different entities, etc.), just like *"for being evil"* and *"for evil"* in Lines #3 and #5 respectively. Another intuitive method—grouping semantically similar full sentences—can relax the constraint of "identical repetition", but is likely to miss cases when only a small part of the sentences have overlapping concepts (e.g., *"Germany"* in Lines #3 and #4).
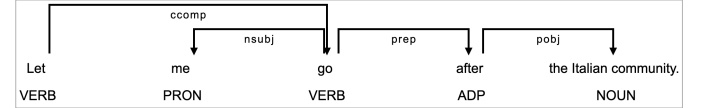

Fig. 6. The dependency tree of Sentence #1 in the Sec. 5.1.3 example.

To capture free-form concepts hidden within full sentences, our grouping method performs two crucial steps: First, **to separate concepts from long sentences**, we induce subphrases by traversing the dependency tree of a given full sentence[5]. For example, with Line 1 parsed into Fig. 6, we get verb phrases like *"go after the Italian community"* as well as noun phrases like *"the Italian community"*. Second, **to merge the rephrasings**, we perform density-based clustering on the induced subphrases based on their semantic similarity: we resolve coreferences between sentences (e.g., *"Their people"* in Line #2 becomes *"Italian people"*)[6]. Then, we transform phrases into feature vectors with a state-of-the-art universal sentence encoder [60] and then compute the cosine similarity in the embedding space. This approach is effective for semantic textual similarity (STS) task (with an accuracy score of 85% on STS Benchmark). Finally, because suphrases recognized through the parsing tree overlap with each other, we reduce redundancy in the extracted repetitions by keeping the segment with the largest possible similarity with its cluster (e.g., in *"go after the Italian community"*, the first two words are considered unnecessary.)

---

5. With the NLP processing library SpaCy: https://spacy.io/
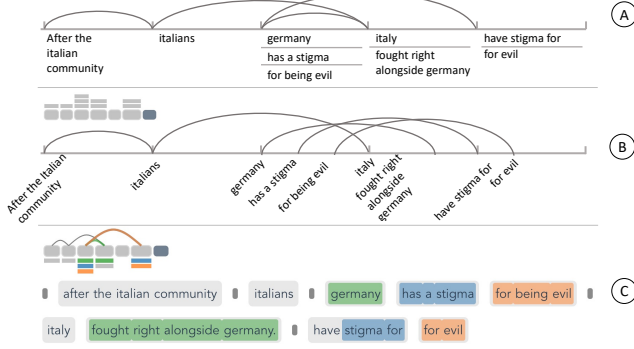6. With https://github.com/huggingface/neuralcoref

Fig. 7. Alternative designs for context linking. Compared to our current design (C), A and B are less space-effective and more cluttered.

Visualizing Contextual Repetitions: We design a context linking graph to display the extracted inter-sentence repetition occurrences. As shown in Fig. 7C, (or a more complete version can be seen in Fig. 9), the graph follows a three-stage design, such that it gradually reveals the concrete context information to the user and traverse from the context-level (**R2**) to the sentence-level (**R3**).

The graph first provides a **context distribution summary**, which shows *how* the sentences are connected to each other through repeated concepts. A rounded gray rectangle represents a sentence, and its horizontal length encodes the sentence length. We highlight the most important punchline with a denser gray color (**R4**). We connect rectangles with arc links if their corresponding sentences share repetitive concepts, and add thin lines below the rectangles to denote the presence of these concept. The combination of the links and the lines help highlight different structures of build-up for humor. For example, Fig. 7C implies that most repetitions occur in the first half of the context, especially the third sentence, with three concepts repeated elsewhere. There is no link between the punchline and the context, suggesting that the punchline is disconnected from the previously mentioned ideas. Then, it presents the concrete **repetitive concepts** to show *what* are used for building up the context, but still omits other details in the related sentences. These concepts are sorted based on the order they occur in the text segment, and each small dark rectangle marks the boundary for a sentence. Finally, the repetitive concept helps locate **detailed sentence** contents and their associated humor features line by line, such that the abstract concepts can be integrated with the complete story.

We also design a set of interactions to enable the traversal of the context summary, repetitive concepts, and detailed sentences. Specifically, when users hover over a rectangle (sentence) of interest in the context summary, all its connections and the corresponding groups of repetitive concepts will be highlighted in different colors. Conversely, as users hover over a phrase in the repetitive concepts, its repeated concepts in other sentences and the corresponding links in the context summary will be highlighted. To facilitate more insights into word usage and verbal delivery, we support a quick reference to the original content in the sentence when users click in context summary or on a specific phrase.

Design alternatives. We discuss the trade-offs of alternative designs for the context summary and repetitive concepts in our iterations. Initially, we attempted to combine the links with concrete concepts. In Fig. 7A, each tick in the

horizontal axis marks the corresponding sentence. Below the tick, repetitive concepts are listed vertically according to the order of their occurrences. While this design does not require separating concepts from the links, this layout could easily exhaust the available horizontal or vertical space when we have a long context or a large number of repeated concepts. It also sacrifices the temporal ordering of the concept occurrence and makes the concept exploration less intuitive. We then tried to place repetitive concepts slantingly along the axis according to their occurrence ordering, and directly link the repeated concepts. Because the notion of "sentence separation" becomes less apparent, we further add a repetition distribution on the top, such that users can count the repeated concepts. The design (Fig. 7B) saves the space and recovers temporal information, but causes visual cluttering issue. Specifically, when the number of concepts increases, linking concepts—instead of their corresponding sentences like in Fig. 7A—induces additional overhead for distinguishing the intertwined links. That said, the concept of overview-to-detail was favored in some preliminary discussions with end-users. Thus, we thought short links among sentence glyphs and concepts overflowed among multiple lines would create the least cognitive load and would be the most space-efficient—which is exactly our design in Fig. 7C.

5.2 Augmented Time Matrix

Besides sentence- and context-level, we design an augmented time matrix that provides an overview of distribution of humor occurrences and the related features of speech content and vocal delivery at the **speech level** (**R1, R2**).

As shown in Fig. 8C, the barcode chart of the time matrix shows the humor distribution. The big gray rectangle shows the whole time matrix from the top to the down. The darker horizontal lines in the time matrix indicate timestamps where the punchlines occurred. Therefore, by definition of the humor snippet (Sec. 4.1), the light gray area between two horizontal lines indicates the context length between punchlines. If the time intervals between punchlines are too small, the horizontal gray lines (i.e., punchlines) are merged into one thick line to reduce visual clutter.

Besides, we organize the humor features, including the word usage, vocal delivery, and key concepts, around the matrix to summarize their distribution for each punchline and across different punchlines. A bar chart is placed at the top to show the total occurrences of humor features in the punchlines, where each bar represents a feature, and the height of the bar indicates the feature frequency. Then, for each punchline, a stacked bar is placed on the left at the same vertical position, where the dark green bar reveals the frequency of textual features, and the light green bar reveals the frequency of audio features. Moreover, colored boxes on the punchlines (dark gray lines) imply the frequencies of humor features. The darker the color, the higher the frequency. To reveal the key concepts for humor snippets, we extract keywords for each snippet using TextRank [61], and place them along the time matrix in temporal order. A horizontal blue bar is overlaid to denote the frequency of the keywords. Users can hover over a feature-of-interest (i.e., a bar at the top or a colored box in the matrix) or a keyword to
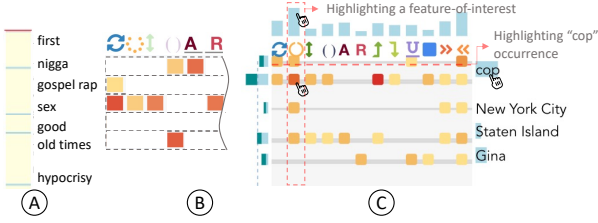
Fig. 8. Alternative designs for speech summary using an annotated barcode chart (A) and a matrix summary of humor features (B). Our current augmented time matrix design (C) combines (A) and (B) to summarize the timing of humor and the distribution of humor features. Users can hover over a feature or a keyword to highlight its occurrences.

highlight its occurrences across the whole speech in the time matrix. When a user clicks a keyword or in the time matrix, the system will highlight the corresponding punchline and its context in the context linking graph.

Design alternatives. Initially, we have considered separating the humor feature summary (Fig. 8B) from the content summary (Fig. 8A). However, the sparse feature matrix takes up a large space and do not provide any context information about the punchline. Particularly, some of our end-users complained that it is hard to figure out where the corresponding punchlines when exploring the feature summary. Thus, in our current design, we integrate the feature summary into the timeline to enhance both the temporal and contextual information for humor features.

5.3 Interactions

Our system supports a rich set of interactions to ease the multi-level exploration of humor **(R5)**.

Details-on-demand through clicking. Once a user clicks a speech of interest in the *control panel*, the *humor exploration* will be updated accordingly. When the user clicks on a keyword or in the augmented time matrix, the corresponding humor snippet will be scrolled to the top in the content exploration, and the corresponding sentence in the context linking graph and the transcript will be highlighted.

Linked scrolling. When users scroll in the content exploration, the time range of the visible humor snippets will be highlighted in the augmented time matrix.

Active query through searching, sorting, and filtering. Users can search a speech or sort speeches according to a specific criterion in the *control panel*. Also, they can apply filters in the *humor focus* to find different styles of punchlines. Then, the corresponding humor snippets will be highlighted in the *humor exploration* view.

6 EVALUATION

We demonstrate how *DeHumor* helps users gain insights into the verbal content and vocal delivery of humor speeches through two case studies and expert interviews. The experts include two humor researchers (*E1*, *E6*) and two communication coaches (*C1*, *C3*). *E1* and *C1* have participated in the design process, while *E6* and *C3* were new to our system before the interviews. Specifically, *E6* holds a master degree in linguistics, and her research focuses on the pragmatics of humor. *C3* has been a communication coach for four years. He is also a stand-up comedian and has performed at famous venues (e.g., Broadway). During

the interviews, the experts used Dehumor to explore two humor datasets, which consist of 157 shows of *Comedy Central Stand-up* and 1,876 *TED Talks*. The cases in Sec. 6.1 and Sec. 6.2 were found by *E1* and *C1*, respectively. All the experts' feedback was collected and reported in Sec. 6.3.

To better illustrate the cases, we highlight the important humor analysis steps: Context relationship analysis for context exploration, Humor context and Punchline for humor description, and Feature analysis for punchline analysis.

6.1 Case Study on Stand-up Comedy

In this case, *E1* used *DeHumor* to explore the "stand-up comedy" dataset and check how comedians effectively set up humor about the funny incidents happened in their lives. In particular, she was interested in the word usage of punchlines and would like to see how it helps create humorous effects. First, *E1* filtered the speeches by keywords "personal experience" in the *control panel* (Fig. 1A1). Then, she felt interested in speeches that frequently involve humorous moments. Thus, she sorted the speeches by the total occurrences of laughter and selected the first speech named "Apparently You Can't Pretend You're a Cop"[7].

Case Context: This case includes three speaker's personal experiences in the selected speech. (1) The speaker talked about his experience at a crime scene. The people there wanted to check if he was a cop. (2) The speaker told a story when he was a teacher. He was once threatened by a student during a fight. (3) Following the previous story, the speaker described that after the fight, both he and the student claimed that they were ready to die.

6.1.1 Overall styles of verbal humor

E1 wondered what the major characteristics of this speaker's humor strategies in his speech are. By observing the height of bars at the top of the augmented time matrix (Fig. 1C), she found that "repetition (🔁)" and "disconnection (◯)" frequently occur in the punchlines. She was curious about *what words the speaker used to create incongruity and how he delivered* (**R1, R4**). As she skimmed through the dark gray lines in the time matrix, she found clusters of punchlines that are close to each other across the whole speech. She wondered *how the speaker set up humor within a short context* (**R2, R3**). To answer the two questions, she adjusted the filters of context length and textual features in the *humor focus* (Fig. 1B). The snippets that satisfied the conditions were highlighted with colored feature statistics in the augmented time matrix (Fig. 1C). Next, she explored them in detail.

6.1.2 Digging into humorous snippets

She found that the first highlighted snippet appeared at the beginning of the speech (Fig. 1C1), where the keyword "cop" was spotted (**R2**). As she hovered over the word, its other occurrences were also marked in dashed red lines in the time matrix, one of which fell into the current snippet of her interest. Then, she clicked the dashed line to locate the corresponding snippet and its context linking graph to see the story development (**R1, R3, R5**). Context relationship analysis Through the links (Fig. 1C4-1) and repeated phrases (*"are you a cop"*, *"'m a cop"*, *"are you*

7. Comedy video url: https://bit.ly/3u1lcZq

Fig. 9. The context linking graph of the first snippet in Fig. 1C2, with Sentence #2, as well as corresponding repeated phrases and their links being highlighted.
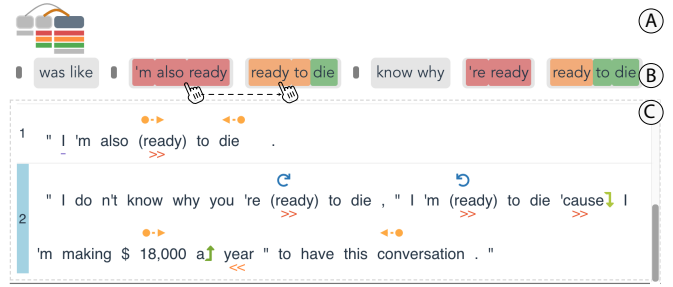

Fig. 10. The context linking graph of the snippet in Fig. 1C3, with the repeated phrases from Sentence #2 and their links being highlighted.

a cop") in the context distribution summary of the graph (Fig. 1C4-2) (*R4*), she guessed that the speaker was having a conversation with someone else about the "cop" identity. She followed the link connections among sentences and observed the corresponding content (Sentences #0 to #2 in Fig. 1C4-3) (*R3*)— Humor context the speaker was asked by the people at a crime scene about whether he was cop. Since he is not an actual cop, he did not want to explicitly make an illegal claim that he was a cop. *E1* navigated to the punchline (Sentence #3 in Fig. 1C4-3) to see how the speaker responded to the question. Punchline She discovered that the speaker quoted a common trope for police in movies and TV (i.e., *"I'm asking the f**king question!"* in Fig. 1C5-3) and misled the people at the crime scene to believe that he was a cop. Feature analysis Specifically, *E1* referred to the feature annotations in the sentence, finding that the speaker raised his voice (⬆) on the first few words in the punchline (*"And"*, *"I"*). Then the speaker paused a little bit (◼) before revealing the essence of the content—"I will ask the f**king (⬆, **( )**) questions". Finally, he strengthened his annoyance by the previous question about his identity (⬆) through a tone particle "okay". *E1* concluded that the "cop" repetitions in the context and his voice modulation in the punchline renders the speaker's annoyance and enhance humorous effect.

Then, *E1* clicked in the second highlighted snippet (*R5*) (Fig. 1C2). Context relationship analysis She noticed that the third rectangle in the context summary (Fig. 9A) has the most bars attached below, suggesting it contains the most repetitive phrases. Then, she clicked on the rectangle and the corresponding repetitions were highlighted. By observing the red rectangles (*"to the ground"*) and purple rectangles (*"going to kill you"*, *"kills"*) of the repeated phrases (Fig. 9B), she assumed there was a big fight. By following the links in the graph from beginning to end and exploring the content (Fig. 9A), Humor context she grasped that the speaker wrestled a student to the ground during a fight. Then, the student threatened that his cousin would come and help him kill the speaker. Punchline & Features The speaker said he also had a cousin (🔄) (Sentence # 4 in Fig. 9C) in response to the student, implying his cousin would also help him and kill the student if the student's cousin killed him (*R4*).

Then, *E1* scrolled down until the third highlighted snippet (Fig. 1C3). Humor context She found that the speaker won the fight with the mentioned student. The student said that he was "ready to die" after losing the fight. Then in the current snippet (Fig. 10), *E1* tracked the colored repeated phrases ("ready to die") (Fig. 10B) (*R3*). She found that

the speaker responded with "I'm also ready to die" and explained in the punchline (Sentence #2 Fig. 10C) (*R4*)— Punchline the speaker felt disappointed about arguing with the naughty student because he was paid extremely low wages at school to deal with such a big trouble maker (i.e., the student). Feature analysis Specifically, the labeled word pair *"18000"* (●··▶) and *"conversation"* (◀·●) in the punchline (Sentence #2 in Fig. 10C) contrasts the low-paid job with the high effort of teaching the student. In addition, the speaker even inserted pauses (◼) and raised his voice (⬆) after *"18000"* to emphasize his complaints about his challenging but low-paid work.

As for takeaways of this exploration process, *E1* concluded that the speaker set up conversation scenarios to narrate his interesting personal experiences. He is good at using contextual repetitions to connect pieces of a story and using words to create incongruity. Moreover, he modulated his voice (e.g., using pauses and increasing volume) to express his emotion and strengthen the humor.

6.2 Case Study on TED Talks

In the second case, the communication coach *C1* explored humor skills in *TED Talk* speeches that are related to "technology". Since lots of his clients come from IT companies, he expected talks on technologies are suitable teaching examples for using humor in speech. In particular, he focused more on the timing of humor and speakers' vocal delivery skills, which were regarded as practical and effective humor skills for students to follow and further improve their speeches. Sorting the speeches by the number of views in descending order, he discovered the most popular speech named *"This is what happens when you reply to spam email"*[8].

Case Context: This case includes two pieces of the speaker's experiences of replying to spam emails. In the speech opening, the speaker introduced that he once received an email from a sender who had a strange name, and described how he replied to the email for fun. To wrap up the speech, the speaker first suggested the audience replying to spam email with a pseudonymous email address.

Originally, *C1* noticed that in the bar code chart (Fig. 11A), the laughter has a dense concentration at the start and end of the timeline, which was often considered as a pattern of strong opening and closing. He clicked the speech to saw how the speaker delivered humor (*R1, R4*) to entertain the audience at the start and the end (*R2*).

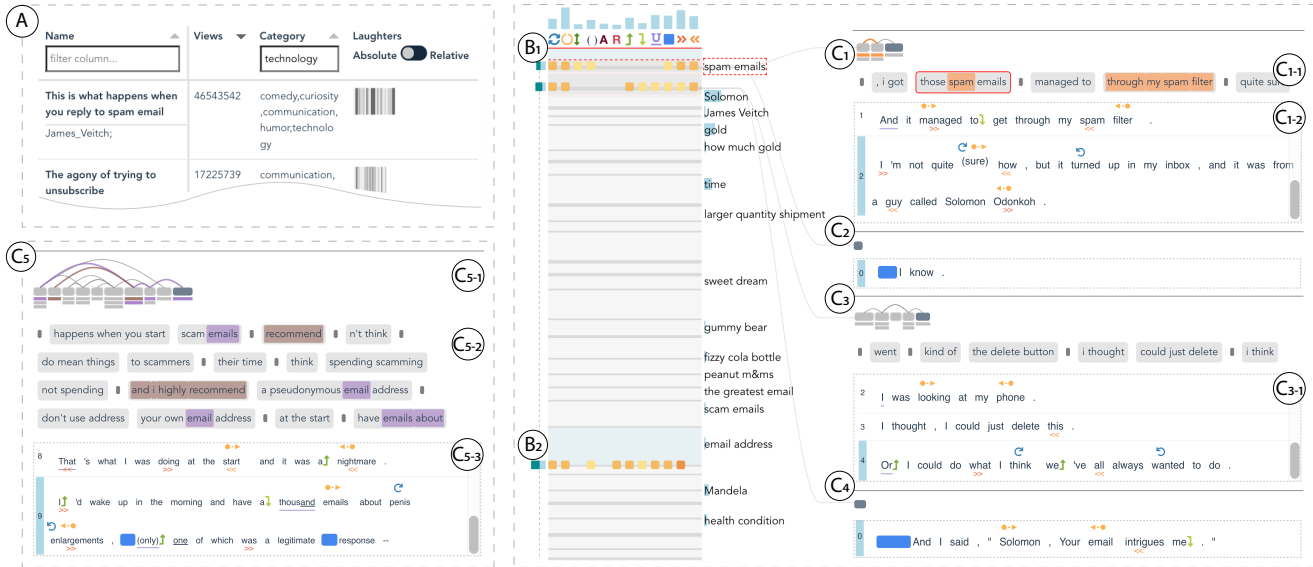8. Ted Talk video url: https://bit.ly/3eFqm6P

Fig. 11. The case study on *TED Talk*. After selecting the speech in (A), the user clicked on "*spam emails*" in the augmented time matrix. Then the snippet in the speech opening (B1) and their context linking graphs (C1-4) are shown. Next, the user used the *humor focus* (Fig. 1B) to find the snippet in the speech closing (B2) with rich humor delivery skills (long dark green bars to the left). Its context linking graph is shown in (C5).

6.2.1 Speech opening

He noticed that "spam emails" appears near the top of augmented time matrix (Fig. 11B1). He clicked the phrase and saw how the speaker introduced it. Context relationship analysis & Humor context From the highlighted phrases *"those spam emails"* and *"through my spam filter"* in the context linking graph (Fig. 11C1-1) (*R2, R4*), he inferred that the speaker received a spam email. Then, he clicked on the highlighted phrases and confirmed his thought after reading the detailed sentences. Punchline In the punchline, *E1* found that the speaker introduced the identity of the spam email sender, who had a very strange name—"Solomon Odonkoh" (Sentence #2 of Fig. 11C1-2). Feature analysis Specifically, *E1* observed that the speaker slowed down his speed on the word "guy (≪)" before revealing the spammer's name. Moreover, the speaker paused (■), and immediately added a short phrase "I know" (Sentence #0 in Fig. 11C2), which triggered another immediate laughter about the stranger for the second time. *C1* commented that the pause and speed variation motivated the audience to think about the spammer's name and identity.

Similarly, *C1* also spotted a pause (■) between the next two snippets (Fig. 11C3, C4) (*R1, R4*), so he explored them accordingly. Humor context He discovered that after introducing the spammer, the speaker also considered deleting the email (Sentence #3 in Fig. 11C3-1). However, he decided not to. Instead, he did what *"we've always wanted to do"* (Sentence #4 in Fig. 11C3-1)—reply to this email. Punchline Then, the speaker shared his funny responses to the email, starting with acknowledgment, *"Solomon, you email intrigues me."* (Sentence #0 in Fig. 11C4). Feature analysis *C1* commented that this was a smart pause (■) at the beginning for helping engage audiences to digest the speaker's previous sentence. Here the audience got a chance to connect *"we've all always wanted to do"* (the bottom of Fig. 11C3) with their desire for replying to spam emails. The pause aroused the audience's interests in the speaker's next move, which enhanced the humorous effect of the speaker's unexpected ac-

knowledgment of the spam email (Sentence #0 in Fig. 11C4).

6.2.2 Speech closing

Then, *C1* wanted to see more snippets at the end, with rich delivery skills, especially with pauses. Thus, *C1* used the *humor focus* to find the highlighted snippet (*R2*) (Fig. 11B2). For the first one, he referred to its context linking graph (Fig. 11C5). Humor context As he tracked the repetition links (Fig. 11C5-1) from the left to the right, he realized that the speaker expressed a positive attitude towards replying to spam email through repeated phrases *"(highly recommend)"* (green rectangles in Fig. 11C5-2) and their sentence contents. The speaker suggested using a *"pseudonymous email address"* to do so and explained the reason in the punchline (Sentence #9 in Fig. 11C5-3). Punchline He once used his own email address. The result is that the mailbox was flooded with *"a thousand"* useless advertisements about *"penis enlargements"*. Among them, he was only able to find one legitimate information that he wanted (Sentence #9 in Fig. 11C5-3). Feature analysis The speaker stressed the reason in the punchline by pauses (■) and vocal stress (U) on keywords such as *"thousand"* and *"only"* (*R1, R4*).

C1 commented that the speaker effectively used pauses to adapt the pace of his presentation to engage the audience, which is considered to be comic timing. He added that pause is crucial to speech delivery, and most students do not realize how powerful it is. He emphasized that this speech is a good example for teaching students how to use pauses to deliver a strong opening and closing in their speeches.

6.3 Expert Interviews

We collected the experts' feedback from the individual interviews with the aforesaid experts (*E1, E6, C1, and C3*). Each interview lasted about one hour and was recorded with the participants' consent. First, we gave the participants a fifteen-minute tutorial outlining the humor features with concrete examples, as well as the visual designs and interactions of *DeHumor*. Then, participants were asked to explore

the speeches introduced in Secs. 6.1 and 6.2 in a think-aloud manner for about forty minutes. For each speech, they were asked to find and explore humor snippets with specific timing and features (**R2**)—snippets that contain (1) words-of-interest at speech opening and (2) humor-features-of-interest at speech closing. Then, within each snippet, they were required to reason about what contributes to the humor. Specifically, they were assigned the following tasks:

1) To examine if our context linking graph effectively highlights related build-ups (**R3**), we asked participants to summarize how the speaker builds up humor.
2) To evaluate our inline feature highlighting (**R1, R4**), we asked participants to identify which part of the punchline contributes the most to laughter in terms of word usage and vocal delivery.
3) To validate extracted features (**R4**), we asked participants to read the original text script, listen to the audio, and voice any features that are out of place.

We then collected post-study feedback on system designs, usefulness, usability, and suggestions for improvements.

6.3.1 Results

Compared with manual browsing and digesting of raw humor speeches, all the participants confirmed the usefulness of computation ability and visualization in the system for assisting humor exploration and analysis.

Concrete humor representations. The participants appreciated that *DeHumor* automatically segments a speech based on the audience laughter. They confirmed it helps them quickly focus on the highlights of humor. And the system offers convenient and user-friendly functions for revealing humor patterns in both speech content and vocal delivery. The context linking graph was generally considered useful for traversing and summarizing humor build-ups. *E1* said, "*The context summary (at the top) helps understand and track the backbones of the story.*" They praised the straight-forward inline annotations of textual and audio features. These annotations help the participants quickly identify the important word use, utterances, and their co-occurrence for creating humor; on the contrary, it is challenging to capture these patterns by watching videos only. *E6* mentioned that "*these annotations, especially the audio feature annotations, successfully guide my attention (to critical parts of the punchline). They vividly capture the delivery patterns within the sentence. I can picture the speaker in front of me giving a speech!*"

Analysis flow. For each speech, the participants explored around nine minutes of the speech content for humor analysis. They confirmed that multi-level humor exploration supported by *DeHumor* aligns well with their general analysis workflow. The most time-consuming task is humor context analysis. But all of the participants could finish it in about three minutes with *DeHumor*. The punchline analysis took them about one minute, and the extracted features were validated within a minute. Finally, the participants could elaborate on what contributes to the humor in a snippet regarding the verbal content and vocal delivery.

During the exploration, the textual incongruity features of punchlines were frequently used to identify the essence of humor. The pitch and pause were found most useful for revealing the key delivery patterns. Also, the participants often relied on the verb and noun phrase repetitions to gain an overview of the humor story development in a snippet.

Usage scenarios. The participants valued the interactive exploration experience with *DeHumor* and were eager to use it in the future. Coaches *C1* and *C3* believed it would be an excellent teaching tool for coaches to show their students how to impress the audience with concrete examples (e.g., where to pause). *E1* confirmed that *DeHumor* provides a corpus with various humor examples and enables rich interactions, which facilitates a systematic study of humor.

Despite the positive feedback above, our participants also identified several limitations of *DeHumor* and provided some suggestions for improvement.

Reliability of feature extraction. Our participants found that the extraction of textual features, especially inter-sentence repetitions and incongruity, contains more errors than the extraction of audio features. For example, they did not find a strong semantic disconnection between "*f**king*" and "*questions*" (Fig. 1C5). However, they thought that the false positives of incongruity usually did not affect humor analysis very much, since they could highlight some critical content words in punchlines for digesting humor context. In contrast, the errors of inter-sentence repetitions might negatively affect their exploration experience. For instance, the phrases ("*i got*", "*managed to*", "*quite sure*") in Fig. 11C1 were not regarded as repetitions, and their presence confused the participants. *E6* thought showing meaningless inter-sentence repetitions is a little distracting. Thus, *E6* was a bit suspicious about the effectiveness of the context summary in the context linking graph, and she tended to directly explore the sentence details.

Learning curve. Though a fifteen-minute tutorial was provided, the participants still needed our guidance to finish some required tasks at the very beginning (i.e., the first ten to fifteen minutes) of their exploration. For example, the participants might not remember all the visual encodings and interactions, and we further explained them. When we illustrated the system designs, the participants found the augmented time matrix was the most complex view. But after several trials, they could successfully utilize it to find snippets of interest for further humor analysis. They claimed it is worth the effort to learn the system features and were willing to use *DeHumor* in their future research or work.

In addition, they have provided us with valuable suggestions. For example, *E1* recommended adding a sentence comparison function to examine the nuances of vocal delivery or word usage in different sentences. *C1* commented that besides texts and voice, visualizing gestures and facial expressions can enhance the analysis of humor techniques.

7 DISCUSSION

Here, we discuss the lessons learned and system generality. We also identify several limitations that need further research in future work, including extending humor features, alleviating algorithm inaccuracy, enhancing system scalability, and enabling personalized humor explorations.

Lessons learned. We learned two important lessons during our system design and evaluation. *1) Social context is important for humor understanding.* During the evaluation, experts pointed out that interpreting humor requires external

knowledge of social context. For example, understanding the humor in Sec. 1 needs to know the "are you a cop" trope in American movies and TV, and the one in Sec. 5.1.3 relates to the WWII history. *2) Compact summary of multimodal features is helpful for multimedia analysis.* Given the multimodality and heterogeneity of humor expression in speeches, we present inline annotations of verbal and vocal features along the text. Experts confirmed the annotations help gain quick insights into speech content and vocal delivery of humor, as well as the relation between them. We believe that the integration of visual representation and multimedia data facilitates intuitive multimedia content understanding.

System generality. While *DeHumor* supports the analysis of speech content and vocal delivery of humor based on audience laughter markers, it can also be extended to evaluate public speaking skills based on other types of audience reaction (e.g., booing). For example, by highlighting the audio features of the speech sentence in a snippet that elicits booing or applause, we can further investigate effective voice modulation skills. In addition, when there is no audience audio, the context linking graph can still be used for text analysis. First, the text can be divided into snippets based on paragraphs or text segmentation algorithms. Then, the context linking graph can visualize contextual repetitions and narrative flows within a text snippet in various forms of literature (e.g., poetry and novel).

Extending humor features. In this work, we derived a set of significant textual and audio humor features to analyze the speech content and its vocal delivery. The proposed features can be enriched and further improved to enhance the understanding of humor. First, as suggested by our experts, it is interesting to explore how features from other modalities contribute to the delivery of humor. For example, facial landmarks [57], [62], [63], and head movements [62] have been mentioned in previous research. How to incorporate features from these modalities in the analysis is a challenging while promising direction. Second, the extraction and visualization of the repeated phrases for the humor build-up can be enhanced. Currently, we focus on inter-sentence repetitions between punchlines for humor context analysis. During the expert interviews, some participants discovered that some interesting repetitions appear across different snippets and incongruous concepts are distributed across different sentences. They wished to explore them. Hence, we plan to extend the contextual repetition algorithm to extract semantically dissimilar phrases between sentences and to highlight repetitions in the whole speech. Additionally, there are other textual features for humor context analysis. For example, funny riddles are used by many comedians and public speakers to entertain and interact with the audience. We plan to extend context-level features to facilitate further study of a humorous story.

Alleviating inaccuracy of feature computation. Through case studies and expert interviews, we showed that the computation and visualization of humor features assisted users in reasoning about humor styles. Inevitably, the imperfection of the algorithms will have harm the effectiveness of *DeHumor*. To alleviate such issues, we will keep improving the feature extraction. Specifically, we plan to label humor features in the sentences and train advanced deep learning models (e.g., BERT [64], GPT-3 [65]) for more accurate computation. Moreover, we will improve the current visualization by encoding model uncertainty. For example, we can give visual hints (e.g., opacity) about the models' accuracy to alert users when the models output features with low confidence scores.

Enhancing system scalability. Our system divides a speech into snippets based on the laughter occurrences. When the transcript is too long with too many punchlines, the exploration of humor snippets will be not so effective. To mitigate such an issue, we can consider merging neighboring humor snippets based on the semantic similarity and temporal proximity. Moreover, with the increasing number of repetitions within a humor snippet, the context linking graph may have visual clutter of links. More advanced interaction techniques are needed to address such issues (e.g., allowing users to interactively reduce and control the visibility of different groups of links).

Enabling personalized humor explorations. *DeHumor* helps users narrow down to a video of interest according to the speech title, speaker, category, and laughter occurrences. In addition, it provides visual cues for users to find humor snippets based on textual and audio features. As suggested by our communication coaches in expert interviews, supporting more complex user queries (e.g., styles of humor feature combinations) would enable a more personalized exploration of humorous content.

Improving system evaluation. The current evaluation is conducted with only four expert users. A long-term study with more domain experts can further validate the usability and effectiveness of *DeHumor*, which is left as future work.

8 CONCLUSION

In this work, we presented *DeHumor*, a visual analytics system for exploring and analyzing humorous snippets in public speaking. We first summarized humor-related features and design requirements based on literature review and user interviews. Then we developed a set of methods for presenting and decomposing multimodal features from a humorous speech. Through case studies on stand-up comedy shows and TED Talks, as well as interviews with domain experts, we demonstrated the usefulness and usability of *DeHumor* in helping users explore and analyze speech content and vocal delivery of humor in speeches.

In future work, we can improve the system usability by supporting humor query and humor style comparison. We plan to integrate more contextual features and features from other modalities (e.g., facial expressions) into the system. We can also apply deep learning models to improve the feature extraction accuracy. Furthermore, we will conduct a long-term study with more experts to further evaluate the system usability and its effectiveness for humor analysis.

REFERENCES

[1] J. Mulholland, *A Handbook of persuasive tactics: A practical language guide*. Routledge, 2003.

[2] P. Wooten, "Humor: an antidote for stress." *Holistic Nursing Practice*, vol. 10, no. 2, pp. 49–56, 1996.

[3] J. Davidson, *The complete guide to public speaking*. Breathing Space Institute, 2003.

[4] V. Raskin, *Semantic mechanisms of humor*. Springer Science & Business Media, 2012, vol. 24.

[5] G. Jefferson, "A technique for inviting laughter and its subsequent acceptance/declination," *Everyday Language: Studies in Ethnomethodology*, pp. 79–96, 1979.

[6] R. Bauman *et al.*, *Story, performance, and event: Contextual studies of oral narrative*. Cambridge University Press, 1986, vol. 10.

[7] A. Schopenhauer, *The world as will and idea*. Library of Alexandria, 1891, vol. 1.

[8] R. Mihalcea and C. Strapparava, "Making computers laugh: Investigations in automatic humor recognition," in *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2005, pp. 531–538.

[9] D. Yang, A. Lavie, C. Dyer, and E. Hovy, "Humor recognition and humor anchor extraction," in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 2367–2376.

[10] R. Mihalcea and C. Strapparava, "Learning to laugh (automatically): Computational models for humor recognition," *Computational Intelligence*, vol. 22, no. 2, pp. 126–142, 2006.

[11] R. Mihalcea and S. Pulman, "Characterizing humour: An exploration of features in humorous texts," in *Proceedings of International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, 2007, pp. 337–347.

[12] R. Zhang and N. Liu, "Recognizing humor on twitter," in *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management*, 2014, pp. 889–898.

[13] J. M. Taylor, "Computational detection of humor: A dream or a nightmare? the ontological semantics approach," in *Proceedings of 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, vol. 3. IEEE, 2009, pp. 429–432.

[14] D. Radev, A. Stent, J. Tetreault, A. Pappu, A. Iliakopoulou, A. Chanfreau, P. de Juan, J. Vallmitjana, A. Jaimes, R. Jha *et al.*, "Humor in collective discourse: Unsupervised funniness detection in the new yorker cartoon caption contest," *arXiv preprint arXiv:1506.08126*, 2015.

[15] L. Pickering, M. Corduas, J. Eisterhold, B. Seifried, A. Eggleston, and S. Attardo, "Prosodic markers of saliency in humorous narratives," *Discourse Processes*, vol. 46, no. 6, pp. 517–540, 2009.

[16] S. Attardo, L. Pickering, and A. Baker, "Prosodic and multimodal markers of humor in conversation," *Pragmatics & Cognition*, vol. 19, no. 2, pp. 224–247, 2011.

[17] S. Attardo and L. Pickering, "Timing in the performance of jokes," *Humor-International Journal of Humor Research*, vol. 24, no. 2, pp. 233–250, 2011.

[18] S. Attardo, L. Pickering, F. Lomotey, and S. Menjo, "Multimodality in conversational humor," *Review of Cognitive Linguistics. Published under the auspices of the Spanish Cognitive Linguistics Association*, vol. 11, no. 2, pp. 402–416, 2013.

[19] C. Kiddon and Y. Brun, "That's what she said: double entendre identification," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers-Volume 2*. Association for Computational Linguistics, 2011, pp. 89–94.

[20] D. Donahue, A. Romanov, and A. Rumshisky, "HumorHawk at SemEval-2017 task 6: Mixing meaning and sound for humor recognition," in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*. Vancouver, Canada: Association for Computational Linguistics, Aug. 2017, pp. 98–102. [Online]. Available: https://www.aclweb.org/anthology/S17-2010

[21] J. M. Taylor and L. J. Mazlack, "Computationally recognizing wordplay in jokes," in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 26, no. 26, 2004.

[22] X. Yan and T. Pedersen, "Duluth at semeval-2017 task 6: Language models in humor detection," *arXiv preprint arXiv:1704.08390*, 2017.

[23] A. Cattle and X. Ma, "Recognizing humour using word associations and humour anchor extraction," in *Proceedings of the 27th International Conference on Computational Linguistics*, 2018, pp. 1849–1858.

[24] V. Ahuja, T. Bali, and N. Singh, "What makes us laugh? investigations into automatic humor classification," in *Proceedings of the Second Workshop on Computational Modeling of People's Opinions, Personality, and Emotions in Social Media*, 2018, pp. 1–9.

[25] A. Purandare and D. Litman, "Humor: Prosody analysis and automatic recognition for f* r* i* e* n* d* s," in *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, 2006, pp. 208–215.

[26] D. Bertero and P. Fung, "A long short-term memory framework for predicting humor in dialogues," in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2016, pp. 130–135.

[27] 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," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 2020, pp. 1–12.

[28] S. Rubin, F. Berthouzoz, G. J. Mysore, and M. Agrawala, "Capture-time feedback for recording scripted narration," in *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*, 2015, pp. 191–199.

[29] A. Watanabe, S. Tomishige, and M. Nakatake, "Speech visualization by integrating features for the hearing impaired," *IEEE Transactions on Speech and Audio Processing*, vol. 8, no. 4, pp. 454–466, 2000.

[30] 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*, vol. 26, no. 1, pp. 927–937, 2019.

[31] M. M. U. Rony, E. Hoque, and N. Hassan, "Claimviz: Visual analytics for identifying and verifying factual claims," in *Proceedings of 2020 IEEE Visualization Conference*. IEEE, 2020, pp. 246–250.

[32] M. El-Assady, V. Gold, C. Acevedo, C. Collins, and D. Keim, "Contovi: Multi-party conversation exploration using topic-space views," in *Computer Graphics Forum*, vol. 35, no. 3. Wiley Online Library, 2016, pp. 431–440.

[33] L. South, M. Schwab, N. Beauchamp, L. Wang, J. Wihbey, and M. A. Borkin, "Debatevis: Visualizing political debates for non-expert users," in *Proceedings of 2020 IEEE Visualization Conference*, 2020, pp. 241–245.

[34] A. Öktem, M. Farrús, and L. Wanner, "Prosograph: a tool for prosody visualisation of large speech corpora," in *Proceedings of the 18th Annual Conference of the International Speech Communication Association (INTERSPEECH 2017); 2017 Aug. 20-24; Stockholm, Sweden. Baixas: ISCA; 2017. p. 809-10*. International Speech Communication Association (ISCA), 2017.

[35] J. Oh, "Text visualization of song lyrics," *Center for Computer Research in Music and Acoustics, Stanford University*, 2010.

[36] R. Patel and W. Furr, "Readn'karaoke: visualizing prosody in children's books for expressive oral reading," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2011, pp. 3203–3206.

[37] N. Cao and W. Cui, "Introduction to text visualization," 2016.

[38] S. Jänicke, A. Geßner, M. Büchler, and G. Scheuermann, "Visualizations for text re-use," in *Proceedings of 2014 International Conference on Information Visualization Theory and Applications (IVAPP)*. IEEE, 2014, pp. 59–70.

[39] R. Vuillemot, T. Clement, C. Plaisant, and A. Kumar, "What's being said near "martha"? exploring name entities in literary text collections," in *Proceedings of 2009 IEEE Symposium on Visual Analytics Science and Technology*. IEEE, 2009, pp. 107–114.

[40] M. Wattenberg, "Arc diagrams: Visualizing structure in strings," in *Proceedings of IEEE Symposium on Information Visualization, 2002. INFOVIS 2002*. IEEE, 2002, pp. 110–116.

[41] A. Don, E. Zheleva, M. Gregory, S. Tarkan, L. Auvil, T. Clement, B. Shneiderman, and C. Plaisant, "Discovering interesting usage patterns in text collections: integrating text mining with visualization," in *Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management*, 2007, pp. 213–222.

[42] I. Subašic and B. Berendt, "Web mining for understanding stories through graph visualisation," in *Proceedings of 2008 Eighth IEEE International Conference on Data Mining*. IEEE, 2008, pp. 570–579.

[43] P. Riehmann, M. Potthast, B. Stein, and B. Froehlich, "Visual

assessment of alleged plagiarism cases," in *Computer Graphics Forum*, vol. 34, no. 3.   Wiley Online Library, 2015, pp. 61–70.

[44] J. Sinclair, *Corpus, concordance, collocation*.   Oxford University Press, 1991.

[45] J. Yuan and M. Liberman, "Speaker identification on the scotus corpus," *Journal of the Acoustical Society of America*, vol. 123, no. 5, p. 3878, 2008.

[46] A. Reyes, P. Rosso, and D. Buscaldi, "From humor recognition to irony detection: The figurative language of social media," *Data & Knowledge Engineering*, vol. 74, pp. 1–12, 2012.

[47] D. Davis, "Communication and humor," *The Primer of Humour Research, Berlin and New York: Mouton de Gruyter*, pp. 543–568, 2008.

[48] S. Castro, L. Chiruzzo, A. Rosá, D. Garat, and G. Moncecchi, "A crowd-annotated spanish corpus for humor analysis," *arXiv preprint arXiv:1710.00477*, 2017.

[49] W. Nash, *The language of humour*.   Routledge, 2014, vol. 16.

[50] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. S. Narayanan, "The interspeech 2010 paralinguistic challenge," in *Proceedings of Eleventh Annual Conference of the International Speech Communication Association*, 2010.

[51] J. H. Goldstein, "Repetition, motive arousal, and humor appreciation." *Journal of Experimental Research in Personality*, 1970.

[52] D. Tannen, *Talking voices: Repetition, dialogue, and imagery in conversational discourse*.   Cambridge University Press, 2007, vol. 26.

[53] M. Halliday, C. M. Matthiessen, and C. Matthiessen, *An introduction to functional grammar*.   Routledge, 2014.

[54] J. R. Martin, *English text: System and structure*.   John Benjamins Publishing, 1992.

[55] J. Morreall, "The philosophy of laughter and humor," 1986.

[56] D. Bertero and P. Fung, "Multimodal deep neural nets for detecting humor in tv sitcoms," in *Proceedings of 2016 IEEE Spoken Language Technology Workshop (SLT)*.   IEEE, 2016, pp. 383–390.

[57] M. K. Hasan, W. Rahman, A. Zadeh, J. Zhong, M. I. Tanveer, L.-P. Morency *et al.*, "Ur-funny: A multimodal language dataset for understanding humor," *arXiv preprint arXiv:1904.06618*, 2019.

[58] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543.

[59] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, 2005, pp. 347–354.

[60] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.   Association for Computational Linguistics, 11 2019. [Online]. Available: http://arxiv.org/abs/1908.10084

[61] R. Mihalcea and P. Tarau, "Textrank: Bringing order into text," in *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, 2004, pp. 404–411.

[62] S. Petridis and M. Pantic, "Is this joke really funny? judging the mirth by audiovisual laughter analysis," in *Proceedings of 2009 IEEE International Conference on Multimedia and Expo*.   IEEE, 2009, pp. 1444–1447.

[63] S. Petridis, B. Martinez, and M. Pantic, "The mahnob laughter database," *Image and Vision Computing*, vol. 31, no. 2, pp. 186–202, 2013.

[64] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pretraining of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[65] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.

Yao Ming is a research scientist at Bloomberg LP. His research focus on visual analytics, explainable machine learning, and natural language processing. He received a Ph.D. in Computer Science from the Hong Kong University of Science and Technology and a B.S. from Tsinghua University. For more details please refer to https://www.myaooo.com

Tongshuang Wu is a fifth year PhD student at the University of Washington, co-advised by Jeffrey Heer and Daniel S. Weld. She received her BE from the Hong Kong University of Science and Technology (HKUST). Her research focuses on helping humans more effectively and systematically evaluate and interact with their models through Human-Computer Interaction (HCI) and Natural Language Processing (NLP). For more details, please refer to https://homes.cs.washington.edu/~wtshuang/.

Haipeng Zeng is currently an assistant professor in School of Intelligent Systems Engineering at the Sun Yat-sen University (SYSU). He obtained a B.S. in Mathematics from Sun Yat-Sen University and a Ph.D. in Computer Science from the Hong Kong University of Science and Technology. His research interests include data visualization, visual analytics, video analysis and machine learning.

Yong Wang is currently an assistant professor in School of Information Systems at Singapore Management University. His research interests include data visualization, visual analytics and explainable machine learning. He obtained his Ph.D. in Computer Science from Hong Kong University of Science and Technology in 2018. He received his B.E. and M.E. from Harbin Institute of Technology and Huazhong University of Science and Technology, respectively. For more details, please refer to http://yong-wang.org.

Xingbo Wang is a Ph.D. candidate in the Department of Computer Science and Engineering at the Hong Kong University of Science and Technology (HKUST). He obtained a B.E. degree from Wuhan University, China in 2018. His research interests include multimedia visualization, interactive machine learning for natural language processing (NLP). For more details, please refer to https://andy-xingbowang.com/.

Huamin Qu is a professor in the Department of Computer Science and Engineering (CSE) at the Hong Kong University of Science and Technology (HKUST) and also the director of the interdisciplinary program office (IPO) of HKUST. He obtained a BS in Mathematics from Xi'an Jiaotong University, China, an MS and a PhD in Computer Science from the Stony Brook University. His main research interests are in visualization and human-computer interaction, with focuses on urban informatics, social network analysis, E-learning, text visualization, and explainable artificial intelligence.