SpeechLens: A Visual Analytics Approach for Exploring Speech Strategies with Textural and Acoustic Features

Conference Paper · February 2019

DOI: 10.1109/BIGCOMP.2019.8679261

CITATIONS
2

READS
44

5 authors, including:

Linping Yuan
The Hong Kong University of Science and Technology
4 PUBLICATIONS **6** CITATIONS

SEE PROFILE

Yuanzhe Chen
Huawei Technologies
23 PUBLICATIONS **505** CITATIONS

SEE PROFILE

SpeechLens: A Visual Analytics Approach for Exploring Speech Strategies with Textural and Acoustic Features

Linping Yuan[1], Yuanzhe Chen[2], Siwei Fu[2], Aoyu Wu[2], and Huamin Qu[2]

[1]Xi'an Jiaotong University
[2]Hong Kong University of Science and Technology
[1]*tuantuan@stu.xjtu.edu.cn*
[2]{*ychench,sfuaa,awuac,huamin*}*@cse.ust.hk*

***Abstract*—Public speaking is an effective way to move, persuade, and inspire. While many guidelines have been presented to teach public speaking skills, they are often based on anecdotal evidence and not customizable. Exploring high-quality speeches such as TED Talks could provide insights to eliminate limitations in existing guidelines. This study aims to explore and identify narration strategies by conducting visual analysis into the textural and acoustic information in public speeches. We present SpeechLens, an interactive visual analytics system to explore large-scale speech dataset with multiple level-of-details. SpeechLens features a novel focus+context design to enable intuitive and smooth analysis. Case studies indicate the effectiveness and usefulness of our approach.**

***Index Terms*—Visual analytics, audio signal processing, public speaking training**

I. INTRODUCTION

While delivering public speech becomes increasingly important, it remains a challenging task for most people since it requires varied skills [6]. One of the major challenges is the difficulties, especially for non-native speakers, to produce an effective and captivating narration of the speech, which has been shown to have an important impact over audience engagement level [7], [9].

A good narration requires speakers to use proper prosody, that is, the melody of speech includes pitch, intensity, speed rate, pause etc., to deliver information expressively. Many systems have been proposed in recent years to help speakers analyze their narration performance. Some work aims at providing instant feedback to speakers during rehearsal [5], [8], [23], [25], and others summarize speaker's performance such as voice modulation to support self-reflection [10], [13]. Recently, Narration Coach [22] was proposed to modify speakers' recordings by re-synthesize technology and allows speakers to hear how they can speak better. Although these systems successfully promote speakers' awareness of their voice status in the presentation, they also have limitations. First, some work requires speakers to repeatedly listen to their recordings and the improved ones, which is inefficient and easily makes users feel frustrated. Second, since a good speech can be delivered with various styles, it is better to allow users to explore and customize the narration style they want to imitate.

The past few years have witnessed more and more digitalized speech data such as TED Talks, which provide rich samples of good narration strategies. While many people consider them as references to better speech narration, it is challenging to identify specific samples of desired narration strategies. First, TED Talks is a large-scale dataset with over 2,900 talks, which makes it impossible for people to search for a suitable sample by watching all the talks. Second, it is still difficult to notice useful strategies even if they only focus on one talk, because they might be overwhelmed by thousands of words speaking at a high speed. Those challenges motivate us to build a public speaking exploring system with a data-driven approach. In this paper, we propose SpeechLens, a visual analytics system that allows users to understand good prosodic patterns in high-quality speech samples, and thus to discover good narration strategies. SpeechLens first extracts the prosodic features of each speech and aligns these features with the script. Then, a three-level hierarchy, i.e., speech-level, sentence-level and word-level, is constructed and visually presented. The system consists of four linked views and rich interactions to facilitate this three-level-of-detail analysis. To verify our method, we conduct case studies using TED Talks data and collect feedback from domain experts.

In summary, the major contributions of this paper are:

- An interactive multi-level visual analytics system that helps speakers explore and understand various prosodic patterns in public speech.
- A novel and scalable visual design based on the focus+context technique to display detailed sentence-level prosodic features.
- Case studies based on real world dataset to evaluate the effectiveness of the proposed method.

II. RELATED WORK

A. Visualization of Prosodic Features

There is a large variety of prosodic features visualization techniques. The most common method is using line charts

978-1-5386-7789-6/19/$31.00 ©2019 IEEE

to encode different feature values along a horizontal time axis [16]. Music Flowgram [12] extended the traditional line chart by encoding different feature values as different visual cues of a line chart such as height, background color and graph color. Instead of mapping features to those channels, Yoshii and Goto proposed Music Thumbnailer [28] to generate thumbnail images representing acoustic features by optimizing top-down visual criteria.

Within the scope of understanding prosodic patterns in speech analysis, it is often necessary to associate prosodic features with scripts. Several systems have been presented to embed prosodic features into script visualization. The most straightforward way to embed prosodic features is drawing a line/bar chart along the script [20], or overlaying a heatmap on the script [17]. Besides, ReadN'Karaoke [21] designed two visualization schemes for multiple prosodic features. One is to manipulate text directly and the other is to augment scripts with overlaid text rendering techniques. Oh [19] further added vertical sparklines with summarized musical features to show the overall structure of songs.

Although these methods can reveal prosodic patterns in public speech, it is tedious for users to explore the whole large-scale dataset. Our method features an overview component, which summarizes prosodic features of each speech, allowing users to effectively identify speeches with desired narration style. Moreover, our focus+context design scales better when analyzing and comparing speech-level prosodic features.

B. Analytics of Public Speaking

Many automated systems have recently been developed to analyze speakers' narration status. Some work generates feedback on various factors by automatically analyzing a user-recorded speech. For example, Presentation Trainer [23] provided users with feedback about voice volume and phonetics pauses. Levis and Pickering [14] utilized basic f0 contour to present voice pitch and teach speakers to use proper intonation in discourse. Recently, Narration Coach [22] is proposed to not only provide feedback to users about their narrations, but also generate an improved version by re-synthesizing the original audio, which iteratively improves users' narrations by informing their weakness. Wu [27] developed a system which enables users to explore presentation techniques in TED Talks.

Some work provides real-time feedback with the help of extra devices. For example, both Presentation Sensei [13] and ROC Speak [10] generated visual summaries from a user recorded video, and the latter system also provided comments from the audience. Logue [8] and Rhema [25] used a Google Glass to inform speakers of their speed rate and volume. AwareMe [5] is a detachable wristband which can be used to increase speakers' awareness of their voice pitch, words per minute and filler words.

Since there is no standard to measure the quality of narration, all the above-mentioned approaches either provide feedback based on heuristics, e.g., do not speak too fast or too low, or define high-quality narration based on users annotation. In this paper, we try to tackle this problem from a data-driven perspective, that is, to provide a visual analytic system to explore and imitate from high-quality public speeches. Our system allows users to identify speech samples according to their desired narration style and understand the characteristics of those good samples, and therefore apply the strategy into their narrations.

III. DESIGN PROCESS

The SpeechLens system aims to help speakers explore a large-scale speech dataset and identify good speech samples with meaningful narration strategies. To inform the design of the visualization system, we need first answer two questions: 1) What prosodic features are insightful for users? 2) How to guide users to useful speech and interpretatively present these prosodic features?

To answer these questions, we first collected potential prosodic features based on a comprehensive literature review. Then, we followed a user-centered design process [18] and collaborated with three domain experts to understand user requirements. All the experts have been engaged in English teaching in universities, and one of the experts has taught a presentation skill course for over 10 years. Based on the literature review and the interviews with experts, we summarized the requirements of our system.

A. Prosodic Features

Among various prosodic features, we selectively identified pitch, volume, and pause, which are consistently considered as important factors that affect speakers' narration quality.

Pitch. The change of pitch value can be used to express the intonation, which is one of the most important prosodic features [3]. Different intonation can deliver different messages. If a speaker uses a small variation of intonation, the resulting speech may sound robotic and the audience can lose focus.

Volume. The variation of volume can help to create various effect during narration. For example, peaks of the volume value are usually used to emphasize a specific word [11].

Pause. A properly pause can help hint the audience that the speaker is about to make an important point, allow the audience to digest previous speech, or simply act as a signal of transition. On the contrary, an unnecessary and unintentional pause may disrupt a sentence.

B. Design Requirements

Based on the interviews with domain experts, we consolidate a set of design requirements in order to effectively derive insights from a large-scale speech dataset.

R1: To support quick identification of speeches with similar narration styles or distinctive speeches. Given a speech dataset, it is important to provide users with an overview that shows groups of speeches sharing similar narration styles or a few speeches as outliers. It gives users a rough idea of the dataset and serves as the entry point of the analysis.

R2: To present speech-level temporal distribution of prosodic features. For one speech, it is necessary to show the evolution of prosodic features. Since speeches may vary

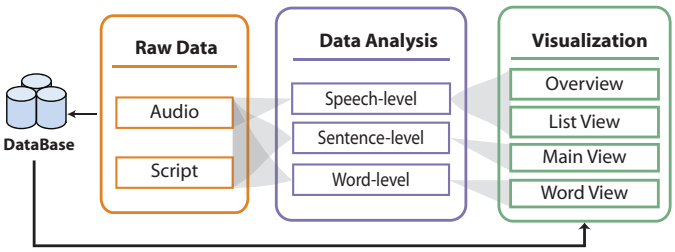

Fig. 1. The system architecture of SpeechLens.

in length and prosodic patterns, the proposed design needs to handle the variance of scales among different speeches.

R3: To present sentence-level distribution of prosodic features. Upon analyzing the prosodic features of a speech, it is helpful to allow users to narrow down to single sentences. The sentence-level design should keep certain prosodic information to keep users being aware of the speech context.

R4: To compare prosodic features of sentences with similar constituent structure. After identifying a sentence with desired prosodic patterns, it is still tedious for users to search for similar one manually. Therefore, the system should provide an automatic method to identify similar sentences.

R5: To summarize prosodic patterns of a specific word or phrase. It is also helpful to summarize prosodic patterns of a selected word/phrase, and hence tell users what kind of narration strategies can be chosen from for that word/phrase.

IV. SpeechLens System

Guided by the aforementioned design requirements, we designed and developed SpeechLens[1], a visual analytics system for interactively exploring narration strategies in large-scale speech dataset. Fig. 1 shows the overview of the system architecture. In the following, we will first describe the data model, and then provide details about the visual design.

A. Data Model

1) Prosodic feature extraction and alignment: In the field of speech analysis and natural language processing, there are many well-established methods related to feature extraction and alignment. Now we describe how to combine several well-known toolkits to construct the data model pipeline.

The first step is to extract prosodic features from audio clips. We adopt a speech analysis toolkit named Praat [2] for feature extraction. The required prosodic features, including pitch and intensity, will be extracted to a form of time series with a predefined sample rate. We chose 0.01 second as the sample rate, which is fine-grained enough for the analysis.

The next step is to align the extracted features with the script. Gentle [1], a robust yet lenient toolkit for aligning speech with text, is used in SpeechLens. After alignment, the start and end timestamp of each word are labeled. We repeat this step for bi-grams, which allows easily drawing prosodic features along the script.

[1]A demo video: https://youtu.be/dtv03qEVFDM

Finally, to enable sentence-level and word-level analysis, we adopt CoreNLP [15] to segment scripts into sentences. Since scripts are already aligned with audio timelines, this step will simultaneously segment the prosodic feature values.

2) Constituent Structure based Sentence Query: As mentioned in Section. III-B, when users identify a useful narration style, the system can benefit users by providing sentences with similar structures. In this way, users can validate and summarize their hypothesis and better understand the narration strategy. However, to the best of our knowledge, existing sentence to sentence querying methods are mostly based on semantics or topics. For example, given an input sentence "I have a dream, a beautiful dream", most existing methods will query sentences talking about "dream". In our scenario, a sentence with a similar structure such as "We have a problem, an environmental problem." is more useful to learn narration styles. Therefore, we propose a constituent structure based similarity (CSS) to measure the distance between two sentences, as stated in Eqn. 1.

$$CSS(S_1, S_2) = \min \sum ||edits(CS_{S_1}, CS_{S_2})|| \quad (1)$$

In Eqn. 1, CS_{S_1} and CS_{S_2} are the constituent sequences of two sentences, $CSS(S_1, S_2)$ is the calculated similarity. To be more specific, for each word/phrase in a sentence, we can easily use a Part-Of-Speech (POS) tagger extractor [26] to extract a tag such as *verb with past tense* or *pronoun*. In this way, a sentence can be transformed into a sequence of POS tags. Then, the CSS can be transformed to the similarity between these two sequences. Therefore, we adopt the Damerau-Levenshtein distance [4], which is a commonly used distance for measuring the similarity of various event sequence data, to finally calculate the CSS.

B. Visual Design

We design the system to fulfill the design requirements discussed in Section. III-B, while following the general design guideline of multiple levels of detail analysis [24]. Fig. 2 shows a screenshot of the user interface. SpeechLens consists of four linked views: the overview which shows the prosodic feature distribution in speech-level, the list view which displays selected speeches with their temporal prosodic feature evolution, the main view supporting sentence-level analysis, and the word view showing the intonation summary of a word.

1) Overview: As shown in Fig. 2(a), we design the overview to illustrate the overall distribution of speeches (**R1**). The overview consists of a scatter plot where each node represents a speech. By default, the x and y-axis represent volume and pitch, respectively. Users can change the axis to encode other attributes, such as average sentence length, sentence count and etc.

2) List View: The list view (Fig. 2(b)) presents the attributes of each speech in a tabular form. The three columns display the speech title, tag and temporal distribution of prosodic features (**R2**). Speeches can be ranked by their word count, sentence count and etc. We visualize the temporal distribution with a

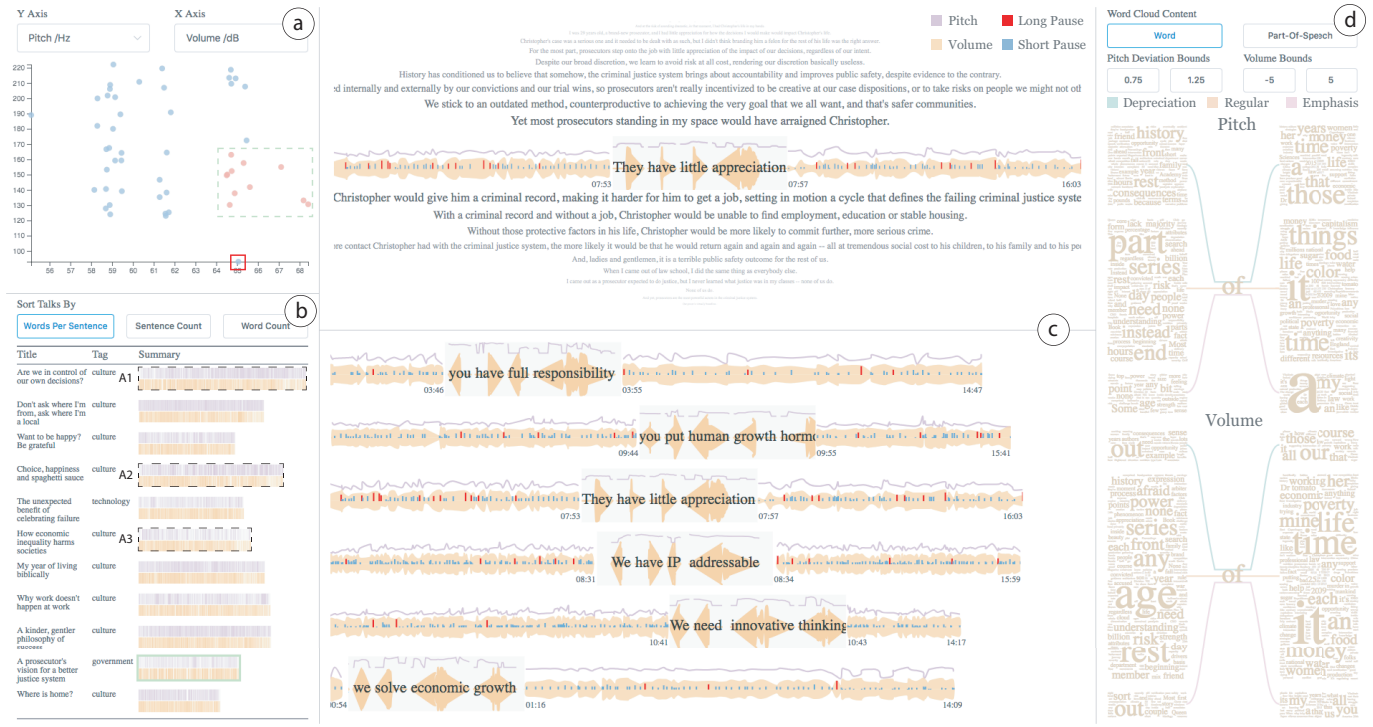

Fig. 2. A screenshot of the proposed visual analytics system for public speech analysis. The system contains an overview (a) which shows the prosodic feature distribution in speech-level, a list view (b) displaying selected speeches with their temporal prosodic feature evolution, a main view (c) supporting sentence-level analysis, and a word view (d) showing the intonation summary of a word.

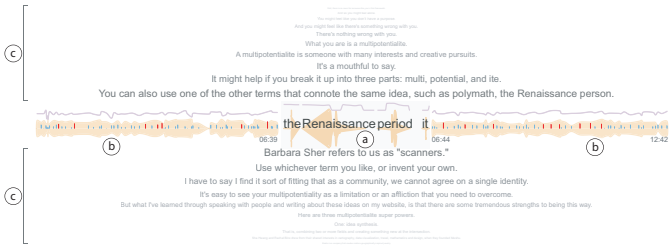

Fig. 3. The focus+context design to display prosodic features of a focused sentence and its context.

space-filling representation. In Fig. 2(b), the top (purple) and the bottom (orange) rows represent the distribution of pitch and volume, respectively. The color opacity encodes the feature value. This color encoding is consistently used in other views.

If users identify a group of speeches they want to further explore, they can click or use lasso selection on the speeches, the list view will be updated to show the filtered results, in this way, users can only focus on the relevant part of the dataset. To further drill down to a lower level analysis, users can click on a speech in the list view, and then the main view will be updated to show the clicked speech.

3) Main View: The main view consists of two parts. The top part shows the prosodic features of a speech with a focused sentence (**R3**). The bottom part visualizes the querying results. Since this is the view displaying both prosodic features and semantics of the script, users can directly learn potentially useful narration strategies by exploring this view. Therefore, the main view is the core view of the SpeechLens system.

Fig. 3 shows the visual design of the top part. We develop a novel focus+context design to preserve the context of a selected sentence while keeping the design scalable to the length of the speech. To be more specific, first we directly place the focused sentence along with a horizontal timeline and overlay its prosodic features on it (Fig. 3(a)). Inspired by [21], we pick the volume chart over the displayed text to encode the volume values, and draw a line chart above the volume chart to present the pitch values. The design rationale is that the width of the volume chart naturally conveys whether the attributes are stronger or weaker, while line chart is more intuitive to present values that become higher or lower. Besides, the duration of pauses between each word is encoded by the blank space between the text, so users can easily observe whether there is a clear pause between words.

Second, we extend the timeline to draw the remained parts (i.e., the context) of the speech before and after the focused sentence (Fig. 3(b)). In this way, the focused sentence has a visual effect as an expanded detail view, which is familiar to analyze according to users' feedback. Users can click on the volume chart to change the focused sentence. To ensure consistent visual encoding, we adopt the same volume chart and line chart to encode volume and pitch, and only change the time scale to fit the context in the screen. Besides, each pause between sentences is encoded by a vertical bar on the timeline. The height of the bar displays the duration of the pause. In this way, users can identify interesting pause patterns (e.g.,

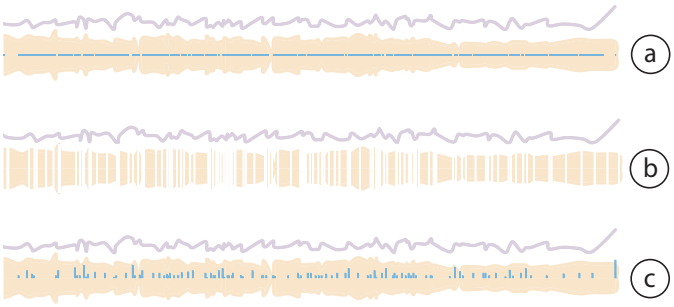

Fig. 4. Design alternatives for encoding pauses. (a) A segmented horizontal timeline. (b) A segmented volume chart. (c) The final design used in SpeechLens.

Fig. 5. Design alternatives for encoding pitch values, including a music node design (top) and the final design used in SpeechLens (bottom).

dense pause usage or unusual long pause) and quickly locate the corresponding sentences.

Last, only showing the text of a single sentence limits users' cognitive ability to understand the content of the speech. Therefore, we also draw the context sentences vertically along the focused sentence (Fig. 3(c)). We decrease the font size and opacity to encode the distance between a sentence and the focused sentence, so a sentence is smaller and lighter if it is further away from the focus.

Iterative design process.

The design of the main view is refined through an iterative process by working with our collaborators. Several design alternatives are considered and implemented during this process.

At first, since we want the visual cues of the focused sentence and the context to be consistent, we design to compress all the scripts to a straight line and use the segments of this line to encode each sentence. Then, the pause between context sentences can be encoded as the blank space between line segments (Fig. 4(a)). Another similar idea is to divide the volume chart into segments and again use the blank space to indicate pause (Fig. 4(b)). However, since the script can have varying length and number of sentences, the available blank space can be limited to clearly show the pause duration. For example, it can be difficult to discriminate a 1 second pause and a 10 second pause. Therefore, we use the bar chart to represent each pause (Fig. 4(c)). In this way, the tall bar indicating unusual long pause can easily catch users' attention.

When encoding the pitch value, we initially consider to use a visual metaphor of music notes (Fig. 5(top)). In this design, we first calculate the average pitch value of each syllable in the

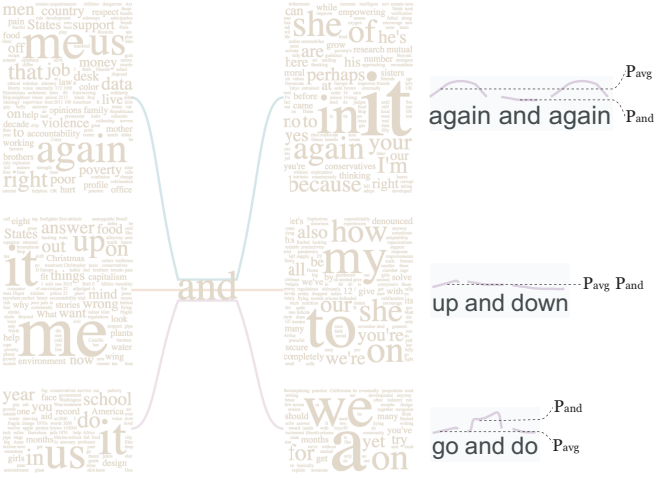

Fig. 6. The extended word cloud design in the word view (left). Occurrences of the selected word are classified into three groups (from top to bottom: depreciation, regular usage and emphasis) according to the prosodic feature values. Examples of the occurrences of each group is shown on the right.

focused sentence, and visually encode it as a music note. We choose this metaphor design because when speakers practice their intonation, they tend to train their pronunciation syllable by syllable. However, after our discussion with collaborators, we finally use the line chart design because: 1) line chart is better in revealing the trend of pitch values, so users can easily observe the intonation of multiple words or a whole sentence. 2) although speakers are familiar with syllable level intonation, we observe that users' cognitive ability can easily match the line chart to each syllable by reading the text. Therefore, we finally choose the design in Fig. 5(bottom).

Similar sentences comparison. Upon users select a focused sentence, the system will take this sentence as input and use the previously described CSS query to search similar sentences. With the query result, the system will also display the prosodic features of these sentences at the bottom part of the main view. To compare and summarize the narration strategies among these sentences, we simply use a side-by-side visual comparison, and encode prosodic features in the same way as the focused sentence (**R4**).

4) Word View: Although the main view can help users find a sentence with the desired narration style, users usually need more samples to understand and generalize their findings. The side-by-side comparison in the main view can provide more sentence samples. Another option is to provide more narration samples for a critical word in the sentence, such as a transition word. The word view is designed for this purpose (**R5**).

To provide a summary of all the sample usages of a given word, firstly, we can easily retrieve all the occurrences of the word. Then, to give users hints about the usage of narration for the word, we classify the intonation of this word into three categories, that is, emphasis, depreciation and regular usage. To illustrate the idea, typical pitch values for each of the category is shown in Fig. 6. According to a previous work [22], both volume value and pitch deviation can help

to detect an emphasized word, in this paper, we also classify the words in a similar but simpler way. To be more specific, we generate two classification results based on the volume value and pitch deviation, respectively. For volume value, given a word, we calculate the average volume value of the antecedent and subsequent words. Then, if the volume value of the selected word is larger or smaller than the average for a threshold λ_1, it is classified as emphasis or depreciation. Otherwise, it is considered as a regular usage. Similarly, we can group all the occurrences based on the pitch deviation with another threshold λ_2. In this paper, we set the two thresholds to 25% and 5dB, respectively, according to [22]. Users can adjust the thresholds through the user interface.

For a selected word, we apply the word cloud to visualize the context information. For each of the three categories, we collect the antecedent and subsequent words of the selected word and generate a word cloud, respectively. As shown in Fig. 6, word clouds are linked with curves, and the color and shape of the curve denote the intonation category. The height of each word cloud represents the number of occurrences of the selected word, allowing users to observe the most commonly used intonation strategy. The word distribution inside word clouds helps users understand the context of the word. By changing the setting, the system allows users to replace all the words with their part-of-speech tags, and generate word clouds based on tag distribution.

V. CASE STUDIES

In this section, we evaluate the effectiveness and usefulness of SpeechLens using case studies. Our users include two domain experts (denoted by E0 and E1) mentioned in Section. III and two graduate students (denoted by S0 and S1).

We collected 51 TED Talk audios and scripts from four topics, i.e., education, government, technology, and culture. Each of them lasted 12 to 18 minutes with 1,536 to 3,111 words and 76 to 263 sentences. We then implemented the system as a web-based application and conducted semi-structured interviews with users. Each interview lasted about one hour. During the interviews, we first introduced the features in SpeechLens, and then allowed users to freely explore the data with the system. Finally, we discussed with users about the insights gained during the exploration, as well as the strengths and weaknesses of SpeechLens. We summarize users' analytic processes as the follows.

A. Speech Overview

First of all, our users wanted to obtain a big picture of the displayed speeches (**R1**). After the data was loaded into SpeechLens, the overview showed the scatter plot with volume and pitch as x and y-axis. S0 noticed that there was a speech with low pitch values, compared with other speeches (marked in red in Fig. 2(a)). He exhibited interests, *"I want to explore this speech because the voice of the speaker may be closer to my deep voice, and maybe I can imitate his narration styles."* E1 changed the x-axis and observed the scatter plots. After changing x-axis to represent average sentence length, E1 mentioned, *"I can easily locate speakers who use complex sentences and those who use short sentences. The difference between these two styles is meaningful to further investigate."*

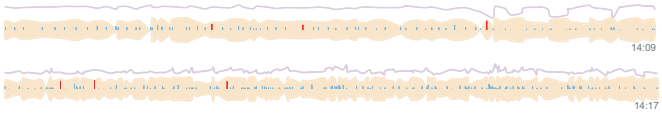

Fig. 7. Speech fingerprints of two TED Talks. One (top) is about *economic growth* which mainly uses explanations. The other (bottom) is about *collaboration between conservatives and liberals* which uses a mixture of explanations, jokes and story-telling.

B. Narration Styles Identified by Prosodic Features

Next, the users wanted to analyze the narration styles at the speech-level (**R2**). S1 was interested in the speeches with relatively high volumes, so he selected them with the lasso tool in the overview. Then he inspected the temporal distribution of each speech in the updated list view. He noticed that most purple rows kept consistent opacity along time while three of them started with high opacity areas (marked as A1-A3 in Fig. 2(b)), indicating low pitch values at the beginning of the corresponding speeches. *"The three speeches may have different narration styles from others"*, he inferred. To find the specific reason, S1 clicked each speech in the list view and then browsed their scripts in the main view. After careful comparison, he found that the three speeches started with explanations while others told stories or asked questions. *"When speakers start their talks with explanations which is usually less emotional, they tend to use low pitch"*, S1 concluded, *"This gives me an insight about starting with explanations. However, I prefer to have a strong opening in my speech, so I would avoid to use it as the opening."*

S1 continued to explore the prosodic features in the main view. With the thought that the context diagram could be considered as a fingerprint of a speech, and could be used to discriminate different narration styles, he paid attention to the diagrams and the pitch curves, volume areas and vertical bars in them. He observed that two speeches had quite different fingerprints. One speech about "economic growth" had a smooth pitch curve and sparse vertical bars (Fig. 7(top)), indicating its small pitch variation and few pauses. In contrast, the fingerprint of the other speech, which was about collaboration between conservatives and liberals, contained a more zigzag pitch curve and much denser vertical bars (Fig. 7(bottom)). After investigating the raw audios and scripts of the two speeches, S1 identified two different narration styles: *"The speaker talking about economic growth doesn't change his pitch a lot, since he just explains the economic phenomenon and uses mostly explanations and long sentences. On the contrary, bigger pitch variation and more pauses are used in the other one, because the speaker is more emotional and uses a mixture of explanation, jokes and story-telling.*

C. Distinct Narration Strategies

To investigate more detailed narration strategies in the sentence-level (**R3**), E0 used the focus+context design to

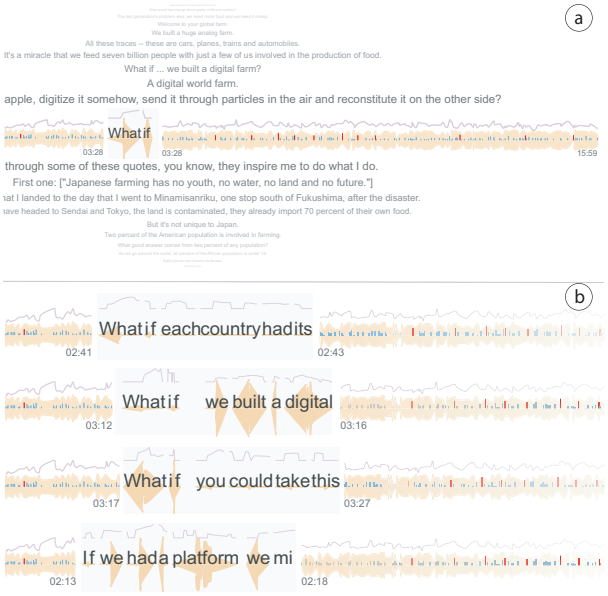

Fig. 8. An example of using strategic silence. (a) A long pause occurs after the phrase *what if*. (b) The querying result shows more similar sentences.

explore a technology speech about a digital farm. E0 browsed its contextual text and found the phrase *what if* occurred several times in different sentences. For further investigation, he clicked on one of them. Then the focused sentence switched to the clicked sentence with its prosodic features (Fig. 8(a)). He noticed the blank space between *what if* and the subsequence words, indicating a long pause. Keeping this pattern in mind, E0 wanted to check the usage of other occurrences of *what if* (**R4**), so he queried this sentence. The screenshot of the querying results is shown in Fig. 8(b).

From the querying results, E0 noticed that there was blank space after the phrase *what if* while no such space appeared after the occurrence of the single word *if*. E0 inferred, *"It seems that the speaker pauses every time he speaks what if. On the contrary, he doesn't pause when he says the single word if. This pattern might suggest a distinct usage of strategic silence after the phrase what if."* After checking the audio, E0 validated his hypothesis. *"The speaker tries to explain his idea of digital farm. He uses the phrase what if and the silence after it to emphasize the impact of the technique and provide space for the audience's imagination. Moreover, the subtle difference between the duration of the pauses prevents the narration from being sounded too mechanical"*, E0 explained, *"This is a good example with such narration strategies. However, I may overlook the pattern when listening to the audio if there are no visualized prosodic features assisting me."*

D. Word Level Intonation Strategies

When exploring the focused sentence, some users clicked on a word, especially commonly used words, to check the summary of its intonation usage (**R5**). Several patterns had been found during this process. Due to the limited space, we report two examples here.

Fig. 9. The intonation of the word *know* (left) and the word *believe* (right).

The first pattern is related to the word *know*. Mostly, words *don't* and *didn't* occur before the word *know* when it is depreciated (Fig. 9(left)). It means that speakers tend to depreciate the word *know* when expressing a negative notion. The second pattern is about the word *believe*. When *believe* is emphasized, it is probably followed by the word *that* (Fig 9(right)). It suggests that the word *believe* is usually emphasized when followed by a clause.

E. General Feedback

Usability. In general, SpeechLens received positive feedback from our users. They mentioned the system was easy to use and appreciated its effectiveness to explore narration strategies. S1 said, *"The visualization of prosodic features makes it easy to discover something special in a speech even when I haven't listened to the speech."* E1 further added, *"With the help of your system, the chance for me to overlook good samples reduces."* However, they also gave us some suggestions. E1 said, *"Currently, I have to visit TED.com and find the raw audios to validate my hypothesis about an interesting prosodic pattern identified in this system. It would be better if the system supports playing the audio."*

Generalizability. During the case study, S0 commented that the system can not only be used for public speaking training, but also for teaching language speaking for non-native speakers, or even general speech analysis. In the future, we plan to interview with experts from various application domains to generalize the usage of SpeechLens.

VI. DISCUSSION

In the following, we report the limitations of our system and discuss potential improvement.

Automatic higher level narration strategy extraction. When compared to directly listening to audios, SpeechLens is more effective on searching for narration styles. However, there are still gaps between users' search intents and the visualized prosodic features. This is because users' intents contain high-level semantics. Therefore, it is useful to embed automatic methods to extract high-level features such as emotion and emphasis from audios and use these features as descriptors to fit users' intents. In the future, we plan to further explore advanced automatic algorithms in the field of audio signal processing to improve the current system.

Word intonation visualization with sentence-level context. Another potential extension of the current system is to summarize the intonation usage of a word within a whole sentence instead of only considering the antecedent and subsequent word. One possible solution is to extend the word cloud with a hierarchical structure, i.e., a tree structure, to aggregate similar words in the same constituent part.

Scalability. Though we only collect 51 TED Talks for the demonstration in the paper, more talks can be easily imported into our system. Our system demonstrates good scalability for long videos due to the use of the focus+context technique. However, when the number of talks increases, the overview may not scale well because of visual clutters. We plan to group videos using clustering algorithms and allow users to select clusters of interest for further exploration.

Evaluation. Some limitations exist in our study design. For example, we involve the same experts during system design and evaluation. However, they are familiar with SpeechLens and may not be able to reveal some potential problems of it. We plan to conduct a more comprehensive user study in the future to evaluate the usability of our system.

VII. Conclusion

In this paper, we have presented SpeechLens, an interactive visual analytics system for exploring and understanding narration strategies in large-scale speech data. SpeechLens displays prosodic features extracted from public speeches in multiple level-of-details. It features a novel scalable focus+context visual design to simultaneously present text and prosodic features. Through in-depth case studies with end users, we demonstrate the effectiveness and usefulness of SpeechLens with real world datasets.

Acknowledgements

The authors would like to thank all the reviewers for their valuable comments. The authors also wish to thank the domain experts and students for their participation in the studies.

References

[1] Gentle. https://lowerquality.com/gentle/.

[2] P. Boersma. Praat, a system for doing phonetics by computer. *Glot international*, 5, 2002.

[3] D. Bolinger and D. L. M. Bolinger. *Intonation and its uses: Melody in grammar and discourse*. Stanford University Press, 1989.

[4] E. Brill and R. C. Moore. An improved error model for noisy channel spelling correction. In *Proceedings of the 2000 Annual Meeting on Association for Computational Linguistics*, pages 286–293. Association for Computational Linguistics, 2000.

[5] M. Bubel, R. Jiang, C. H. Lee, W. Shi, and A. Tse. Awareme: Addressing fear of public speech through awareness. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, CHI EA '16, pages 68–73. ACM, 2016.

[6] M. Chollet, H. Prendinger, and S. Scherer. Native vs. non-native language fluency implications on multimodal interaction for interpersonal skills training. In *Proceedings of the 2016 ACM International Conference on Multimodal Interaction*, ICMI '16, pages 386–393. ACM, 2016.

[7] K. Curtis, G. J. Jones, and N. Campbell. Effects of good speaking techniques on audience engagement. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, pages 35–42. ACM, 2015.

[8] I. Damian, C. S. S. Tan, T. Baur, J. Schöning, K. Luyten, and E. André. Augmenting social interactions: Realtime behavioural feedback using social signal processing techniques. In *Proceedings of the 2015 Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 565–574. ACM, 2015.

[9] J. A. DeVito. *The essential elements of public speaking*. Allyn & Bacon, 2005.

[10] M. Fung, Y. Jin, R. Zhao, and M. E. Hoque. Roc speak: Semi-automated personalized feedback on nonverbal behavior from recorded videos. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, pages 1167–1178. ACM, 2015.

[11] D. Goldberg. The voice over technique guidebook with industry overview. *Edge Studio*, 2010.

[12] D. Jeong and J. Nam. Visualizing music in its entirety using acoustic features: Music flowgram. In *Proceedings of the International Conference on Technologies for Music Notation and Representation, Anglia Ruskin University. Anglia Ruskin University*, pages 25–32, 2016.

[13] K. Kurihara, M. Goto, J. Ogata, Y. Matsusaka, and T. Igarashi. Presentation sensei: A presentation training system using speech and image processing. In *Proceedings of the 2007 International Conference on Multimodal Interfaces*, ICMI '07, pages 358–365. ACM, 2007.

[14] J. Levis and L. Pickering. Teaching intonation in discourse using speech visualization technology. *System*, 32(4):505 – 524, 2004. Incorporating Multimedia Capability in the Reporting of Applied Linguistics Research.

[15] C. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. Bethard, and D. McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of the 2014 Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60, 2014.

[16] P. Mertens. The prosogram: Semi-automatic transcription of prosody based on a tonal perception model. In *Speech Prosody 2004, International Conference*, 2004.

[17] L. C. Milton and C. Y. Lu. Versevis : Visualization of spoken features in poetry. 2011.

[18] T. Munzner. A nested process model for visualization design and validation. *IEEE Transactions on Visualization & Computer Graphics*, (6):921–928, 2009.

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

[20] A. Öktem, M. Farrús, and L. Wanner. Prosograph: a tool for prosody visualisation of large speech corpora. In *Proceedings of the 2017 Annual Conference of the International Speech Communication Association*, ISCA '17, 2017.

[21] 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*, CHI '11, pages 3203–3206. ACM, 2011.

[22] S. Rubin, F. Berthouzoz, G. J. Mysore, and M. Agrawala. Capture-time feedback for recording scripted narration. In *Proceedings of the 2015 Annual ACM Symposium on User Interface Software & Technology*, UIST '15, pages 191–199. ACM, 2015.

[23] J. Schneider, D. Börner, P. van Rosmalen, and M. Specht. Presentation trainer, your public speaking multimodal coach. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ICMI '15, pages 539–546. ACM, 2015.

[24] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE Symposium on Visual Languages*, pages 336–343. IEEE, 1996.

[25] M. I. Tanveer, E. Lin, and M. E. Hoque. Rhema: A real-time in-situ intelligent interface to help people with public speaking. In *Proceedings of the 2015 International Conference on Intelligent User Interfaces*, IUI '15, pages 286–295. ACM, 2015.

[26] K. Toutanova, D. Klein, C. D. Manning, and Y. Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, pages 173–180. Association for Computational Linguistics, 2003.

[27] A. Wu and H. Qu. Multimodal analysis of video collections: Visual exploration of presentation techniques in ted talks. *IEEE Transactions on Visualization and Computer Graphics*, in press.

[28] K. Yoshii and M. Goto. Music thumbnailer: Visualizing musical pieces in thumbnail images based on acoustic features. In *International Society for Music Information Retrieval*, pages 211–216, 2008.