VoiceCoach: Interactive Evidence-based Training for Voice Modulation Skills in Public Speaking

Xingbo Wang, Haipeng Zeng, Yong Wang*, Aoyu Wu, Zhida Sun, Xiaojuan Ma, Huamin Qu
Department of Computer Science and Engineering,
The Hong Kong University of Science and Technology, Hong Kong, China
{xingbo.wang, hzengac, ywangct, awuac, zhida.sun}@connect.ust.hk, {mxj, huamin}@cse.ust.hk

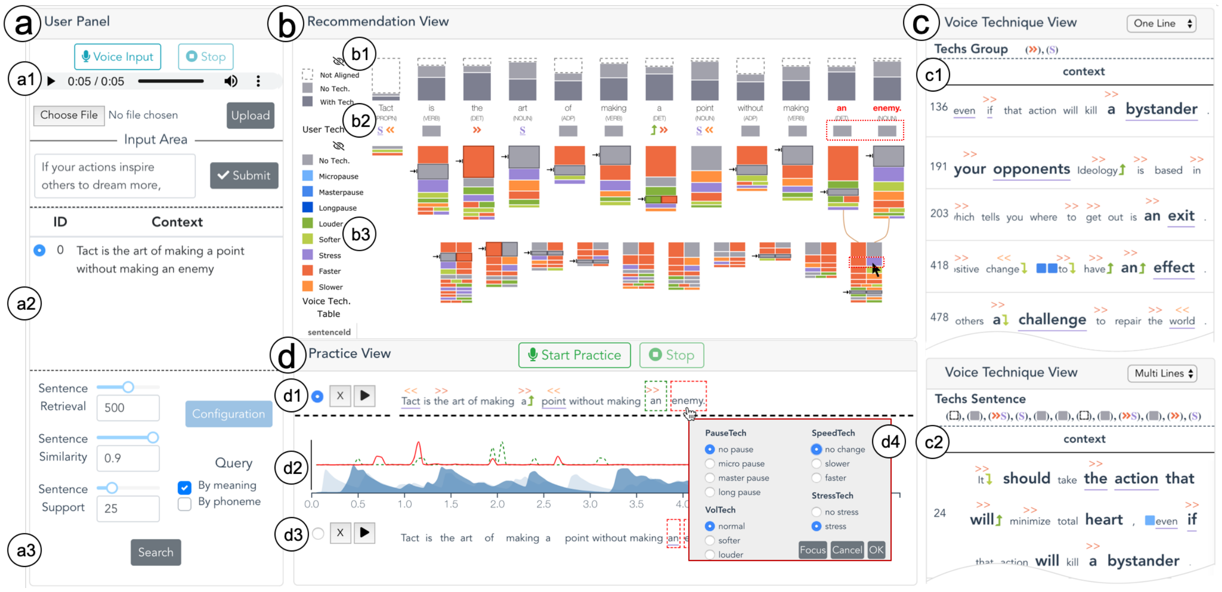


Figure 1: The user interface of *VoiceCoach*: (a) The user panel allows users to submit a query sentence via audio or text input. (b) The recommendation view presents different levels of recommendation results of modulation combination. (c) The voice technique view enables users to quickly locate and compare the contexts of a specific voice modulation skill in either one-line mode or multi-line mode. (d) The practice view provides users with real-time and quantitative visual feedback to iteratively practice voice modulation skills.

ABSTRACT

The modulation of voice properties, such as pitch, volume, and speed, is crucial for delivering a successful public speech. However, it is challenging to master different voice modulation skills. Though many guidelines are available, they are often not practical enough to be applied in different public speaking situations, especially for novice speakers. We present *VoiceCoach*, an interactive evidence-based approach to facilitate the effective training of voice modulation skills. Specifically, we have analyzed the voice modulation skills from 2623 high-quality speeches (i.e., *TED Talks*) and use them as the benchmark dataset. Given a voice input, *VoiceCoach* automatically recommends good voice modulation examples from the dataset based on the similarity of both sentence structures and voice modulation skills. Immediate and quantitative visual feedback is provided to guide further improvement. The expert interviews and the user study provide support for the effectiveness and usability of *VoiceCoach*.

Author Keywords

Voice modulation; evidence-based training; data visualization; public speaking.

CCS Concepts

•**Human-centered computing** → **Human computer interaction (HCI); Visualization; User interface design;**

*Corresponding author.

CHI'20, April 25–30, 2020, Honolulu, HI, USA

ACM ISBN 978-1-4503-6708-0/20/04...$15.00
DOI: https://doi.org/10.1145/3313831.3376726

INTRODUCTION

Public speaking is one of the most important interpersonal skills for both our everyday lives and careers. When delivering a public speech, voice is the primary channel for the speaker to communicate with the audience [12]. Therefore, voice modulation, the manipulation of vocal properties, has a great influence on audience engagement and the delivery of presentations [15]. Many studies [6, 15, 20] have identified key elements for voice modulation including pitch, volume, pause, and speed. For example, increasing the speech speed can convey excitement, while slowing down and using appropriate pauses gives audiences time to reflect on the speaker's words and form personal connections with the content. Higher volume leads to a vocal emphasis. A suitable repetition in a similar voice pitch promotes clarity and enhances effectiveness in making key points. All these voice modulation skills have been proved critical to successful public speaking.

However, it is challenging to master and apply various voice modulation skills in public speaking. Speakers can train themselves by following the general guidelines from the books on public speaking. However, this method suffers from the lack of immediate feedback, because it is often difficult for novice speakers to evaluate their voice accurately. Another possible method is to join a training programs and seek help from professional coaches. However, the feedback from coaches could be subjected to their personal preferences and be inconsistent. There lacks a quantitative method for evaluating speakers' performance and improvements in voice modulation skills. Moreover, it remains unclear about how to combine different public speaking skills and adapt them to different speaking contents and presentation scenarios.

Several prior studies [33, 29, 25] have proposed computer-aided user interfaces to assist in voice modulation training by providing automated feedback on vocal properties such as volume. However, such feedback is determined by constant predefined thresholds, therefore failing to adapt to different speech contexts such as the content and the presentation purpose. Besides, those systems do not provide concrete examples, which could make the learning process less effective.

To address the aforementioned challenges, we aim to develop an interactive system to support effective evidence-based training of voice modulation skills. We work closely with experienced public speaking coaches from an international training company for the past eight months to identify the challenges and detailed requirements. Our resulting system, *VoiceCoach*, helps speakers improve their voice modulation skills in four dimensions, i.e., *pause*, *volume*, *pitch*, and *speed*. We first process 2,623 TED talk videos into over 300,000 audio segments based on sentences and build the benchmark of "good" examples for voice modulation. Then, we propose an effective recommendation approach to retrieve speaking examples for the speakers to explore and learn what can be improved in their voice. The recommended examples are ranked by their similarity with the user input in terms of both speech content (i.e., text) and modulation patterns, facilitating the usage of the most appropriate voice modulation skills for the input sentence(s). This evidence-based training offers novice speakers **concrete and personalized guidelines** about what voice modulation skills can be used. Furthermore, during the practice of voice modulation skills, *VoiceCoach* provides **on-the-fly feedback** on the speaker's performance in terms of mastering the desired voice modulation skills, which is achieved by recognizing the differences between voice modulation skills in the speaker's speaking and the desired ones. Considering that novice speakers may not necessarily have a background in visualization or even computer science, we propose **straightforward visual designs** and make them as intuitive as possible to convey the concrete guidelines and on-the-fly feedback for the training of voice modulation skills.

Our major contributions can be summarized as follows:

- We present an interactive visual system, *VoiceCoach* , to assist in the effective training of voice modulation skills in public speaking. *VoiceCoach* can recommend concrete training examples from a TED talk database based on the semantic and modulation similarities, and supports on-the-fly quantitative feedback to guide the further improvement.
- We conduct expert interviews with professional coaches and a user study with university students, which provide support for the effectiveness and usability of *VoiceCoach* in facilitating the self-training of voice modulation skills.

RELATED WORK

Our work is related to voice modulation, training systems for public speaking, and visualization of audio features.

Voice Modulation

Voice modulation refers to the manipulation of properties of voice [24], including pitch, volume, speed, etc. Researchers have conducted extensive studies on the voice modulation skills in the domain of public speaking, attempting to identify vocal techniques that contribute to a successful speech. Strangert [30] analyzed speech behaviours of news announcers and politicians and summarized the characteristics of "good" speakers, where it was identified that pauses, changes of speed and dynamics of prosody made the speech efficient. Tsai [35] compared vocal characteristics of TED talkers with that of university professors and found out that TED speakers speak at a more consistent speed and with a deeper voice. Rosenberg et al. [10] examined the lexical and acoustic properties of charismatic speech. These studies shed lights on the effective vocal skills for public speaking. However, how to help users quickly and effectively train themselves to master these voice modulation skills still requires further exploration.

Training Systems for Public Speaking

Researchers in the HCI community have proposed several speech training systems, which offer automated feedback on users' speech quality. Many systems [4, 33, 34] evaluate speech quality by measuring vocal characteristics such as pitch, speech rate, and loudness. Their approaches quantify the quality by predefined thresholds regardless of sentences and contexts, which offers insufficient support for deliberate practice of particular sentences. To address this problem, Narration Coach el al. [25] assisted users in recording a script by providing feedback on whether users satisfy voice modulation requirements that are specific to each sentence. However, it requires users to specify those requirements such as spoken emphasis, which could be tedious and particularly difficult for novice users. Therefore, our work studies how to automatically generate voice modulation strategies given an input script. Specifically, we analyze voice modulation strategies from 2,623 high-quality speeches (i.e., *TED Talks*) which are used as the benchmark to recommend strategies.

Another key contributing factor of training outcomes is the feedback strategy. A large body of works has focused on providing in-situ feedback [4, 5, 27, 28, 33]. While such timely feedback is effective for immediate self-correction, long-term retention has been shown to be associated with intermittent feedback [26]. Thus, another line of research proposes interactive systems for analyzing offline feedback to enhance self-reflection [11, 14, 32, 37]. However, those systems only utilize simple charts with limited interaction support, and therefore are insufficient in helping users compare their performance and practice deliberately. Our work combines and extends both strategies by proposing a novel interactive visualization system to convey on-the-fly feedback, and by providing rich interactions to assist in analyzing performance in comparison with recommended examples in an iterative manner.

Visualization of Audio Features

Visualization is an intuitive and effective way of revealing patterns in audio. Much research has focused on developing visualization techniques to represent audio features. One of the most common methods is to use line charts to display temporal changes of feature values [16, 36]. Music flowgram [13] extends line charts by introducing more visual elements such as color and height to encode features. Some works adopt matrix-like [7] or graph-based [19] visualization to describe the structural information of audio. Others utilize metaphors such as clocks [1] and geographical maps [22, 18].

Considering the scenario of speech analysis, audio is often associated with words. Therefore, many visual systems have been developed to explore the relationship between audios and texts. The idea is to overlay audio features along with the scripts. Prosograph [21] horizontally aligns all the words with their corresponding prosodic features, enabling easy exploration of speech corpus. VerseVis [17] draws a filled-line graph, whose height encodes phonemes and color encodes accents. Patel and Furr [23] explores two ways of combining prosodic features with texts: one is to directly control properties of text, using horizontal position, horizontal placement and level of greyness to indicate duration, pitch and intensity respectively. The other is to augment text information by overlaying corresponding prosodic contours.

Although all these works ease the process of tracking temporal changes in the audio features, it requires extra time to both identify the repetitive patterns in lines of scripts and compare structural similarities of features. In comparison, our design gives a quick overview of frequent patterns in the audio collections by displaying technique combinations of varying lengths in a hierarchical order. Furthermore, we convert continuous audio features into compact and intuitive glyphs to facilitate quick analysis of the similarity between lines of words.

DESIGN PROCESS AND REQUIREMENT ANALYSIS

VoiceCoach aims to help novice speakers understand, practice and improve their voice modulation skills. To understand the current practice and challenges of the training process, our design process started with an eight-hour training session offered by our industry collaborator, an international communication and leadership training company. During the training, we conducted contextual inquiries to collect information about the training process and difficulties encountered by trainees, which motivated the initial design of our system. In the later stages, we adopted an iterative development approach by carrying out bi-weekly meetings with four domain experts (*E1-E4*) for eight months. The experts are professional coaches from our industry collaborator, who all have at least six years' experience in the training of professional public speaking. During the meetings, we collected experts' feedback on our early prototypes and updated the system design. Similar to the system design process of prior research [2], these experts serve as proxies to our target population in the requirement analysis and system design of *VoiceCoach*. Specifically, the experts' expertise in public speaking helps us gain a deeper understanding of voice modulation skills. Their experience of public speaking training makes them better aware of the difficulties that novice speakers may encounter in improving their modulation skills. Also, they have deep insights into the limitations of traditional methods for training voice modulation skills.

The design requirements are formulated throughout the eight months and we summarize them as follows:

R1. Inform speakers of their voice modulation. All experts emphasized the importance of providing feedback to speakers on their performance of communication training, which is considered as the basis for improvements. For example, *E3* pointed out that the trainees usually overestimate the time they have paused when practicing the *three-second pause* strategy, but underestimate their volume or pitch. Therefore, it is important to inform speakers of their usage of voice modulation skills.

R2. Provide hints and evidence to guide potential improvements in speakers' voice. According to our expert interviews, another major challenge for novice speakers is how to practice and improve their voice modulation skills. For instance, *E4* said *"Guidance is really important to novice speakers. They usually don't know how and when they need to use voice modulation skills."* Thus, the system should help users quickly identify the issues or problems in their speech and further provide hints to guide their subsequent training based on their performance and preference.

R3. Illustrate the evidence with concrete examples. Our experts commented that they usually provide high-level tips such as *"vary your tone more"*, *"pause longer"* during the training session due to limited time. Such tips, however, could be abstract and difficult for trainees to understand and apply correctly. The system should provide concrete illustrations of voice modulation to promote efficient "learning-by-examples".

R4. Enable on-the-fly feedback on speakers' vocal performance. During the iterative development process, we have found that users sometimes fail to make correct adjustments to their voice modulation when speaking the script, as it is often difficult to memorize all the details of their previous practices. On-the-fly feedback could guide adjustments in a timely manner, making the practice more efficient and effective.

R5. Promote deliberate and iterative practice. We have also observed that speakers could only focus on a few aspects during each practice. *E4* commented *“Most people can’t apply all types of modulation skills into one sentence and it is good enough to have two or three voice modulation skills on meaningful words or phrases.” E1* said *“We cannot expect people to master voice modulation at the first try.”* Therefore, the system should enable and encourage them to focus on specific types of voice modulation skills in an iterative manner, helping speakers practice and improve deliberately.

VOICECOACH

According to the aforementioned system requirements, we further design and implement *VoiceCoach* (Figure 1), an interactive system for exploring and practicing voice modulation skills. The system architecture (Figure 2) consists of four major modules, i.e., data preparation, speech analysis, recommendation engine, and user interface. The data preparation module creates the benchmark for voice modulation training. The speech analysis module analyzes modulation skills in users’ audio input. The recommendation module retrieves good learning examples based on the input. The user interface module enables effective exploration and comparison of voice modulation skills in the retrieval results, and provides real-time quantitative feedback on users’ performance.

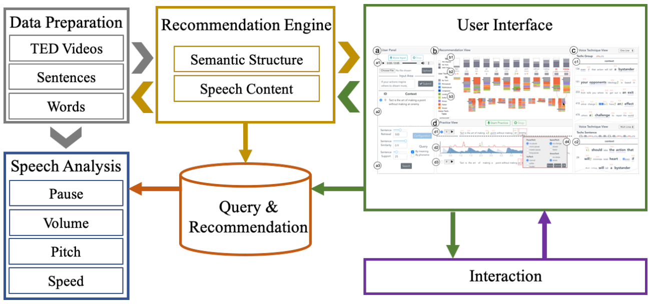


Figure 2: The system architecture of *VoiceCoach*, which is comprised of four major modules, i.e., data preparation, speech analysis, recommendation engine, and user interface.

Data Preparation

The data preparation module aims to create a database of high-quality speeches that are used as the benchmark for training. We choose TED Talks because they are widely considered as the pinnacle of public speaking in terms of high-quality speech content and presentation skills [8]. According to the official TED organizer guide [1], the recording equipment is carefully set and tested to ensure a constantly good audio quality. The invited speakers have diverse professional backgrounds (e.g., entrepreneurs, educators) and the speeches cover over 400 topics. Prior researches [31, 35] have also used them as the benchmark for audio analysis of presentation styles.

We collect videos from the TED Talk website[2] published until June 2019. Then, audio clips of the videos are converted to scripts using the Amazon Transcribe API[3]. A summary of the dataset is shown in Table 1. For each talk, the transcribed texts are split into sentences at periods, exclamation marks or question marks. Each sentence contains all the spoken words together with their start and end time records.

Table 1: Dataset Properties.

Property	**Quantity**
Total number of talks	2,623
Total length of all talks	585.85 hours
Average duration of a talk	13.4 minutes
Total word count	5,350,391
Total sentence count	334,692
Average words of a sentence	15.99
Total topic categories	430
Total speakers’ occupations	447

Speech Analysis

After a user uploads his/her audio input, the speech analysis module will process the audio and detect the employed modulation skills in terms of four vocal properties (i.e., pause, volume, pitch, and speed):

Pause: We focus on intentional pause other than unnecessary interruptions. We calculate it by measuring the interval between two words, which are classified according to coaches’ training specifications - [0.5s, 1s): “brief pause”, [1s, 2.5s): “master pause”, and [2.5s, ∞): “long pause”.

Volume: We compute the average volume for each word, as well as the average and the standard deviation (SD) for each sentence. Then, we label words that are louder ($>$ 1.1 times or $>$ 1 SD) than the sentence as “louder” and softer ($<$ 0.67 times or < -1 SD) than the sentence as “softer”.

Pitch: A higher pitch relates to a vocal stress. Similar to volume calculation, we track the pitch contours to find peak values. Specifically, we label words that are higher pitched ($>$ 1.25 times) or have more pitch variation ($>$ 1 SD) as “stress”.

Speed: We consider two variations of speed (i.e., faster and slower). We compute the Syllables Per Minute (SPM) for each word, as well as the average and standard deviation of SPM for each sentence. Then, we label those that are faster ($>$ 1.5 times) or have more variation ($>$ 1 SD) as “faster” and slower ($<$ 0.67 times) or have more variation ($<$ -1 SD) as “slower”.

We set the above default thresholds empirically together with our experts, and also allow users to change them in the user panel (Figure 1(a3)) to enable interactive customization by users when necessary.

Recommendation Engine

The recommendation module retrieves TED speech examples from the TED dataset by considering both the semantic

[1] https://www.ted.com/participate/organize-a-local-tedx-event/tedx-organizer-guide

[2] https://www.ted.com/talks

[3] https://aws.amazon.com/transcribe/

structure of speech content and the voice modulation skills employed in the user input. It consists of three phases. The first phase is to search semantically relevant sentences in the database. We leverage the state-of-the-art sentence encoder method [3] to embed sentences into feature vectors that preserve the semantic information. Then, the recommendation module finds examples that are close to the query based on cosine similarities in the high-dimensional embedding space. To speed up the search, we leverage Annoy [4], which is one the most popularly-used nearest neighbor search libraries and has been used in the recommendation engine in Spotify[5]. Hence, we retrieve a set of semantically relevant sentences from the dataset. The second phase is to align the sentence of user input with the retrieved sentences based on structural information. The retrieval results from the first phase will be aligned with the input query based on part-of-speech features to facilitate the comparison of sentence structures and voice modulation skills employed in the corresponding sentences. The third phase is to search frequent modulation combinations based on aligned words in the retrieved sentences. At this phase, the recommendation module recommends the usage of voice modulation based on n-grams, which incorporates different lengths of word contexts. It constructs a FP tree [9] on the structurally aligned technique sequences of retrieved examples, and finds frequent voice modulation combinations in the tree. A high support threshold value decreases the generated combinations, while a low one reserves more unusual combinations. The default value is 0.05, which can be interactively adjusted by users in the user panel (Figure 1(a3)).

User Interface

To make the voice modulation training more user-friendly, we design an interactive visual analytics system called *Voice-Coach* (Figure 1) with four coordinated views, including (a) user panel, (b) recommendation view, (c) voice technique view, and (d) practice view. The voice modulations are visually encoded by both colors and glyphs, as shown in Table 2.

Table 2: Glyph Encoding of Voice Modulations

Property	Modulation	Glyph
Speed	Faster	»
	Slower	«
Volume	Louder	↑
	Softer	↓
Pause	Brief pause	■
	Master pause	■ ■
	Long Pause	■ ■ ■
Pitch	Stress	S
None	No Tech.	■

User Panel

The user panel accepts different types of user input including audio streaming, audio files, and texts in Figure 1(a1). Then, it presents the resulting sentences in a table in Figure 1(a2). The user can adjust the parameters of both the example retrieval algorithms and the voice modulation detection approach at the bottom (Figure 1(a3)).

Recommendation View

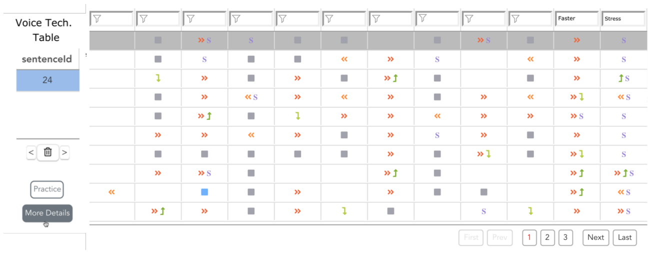
After the system analyzes the user input and retrieves "good" examples from the benchmark dataset, the recommendation view (Figure 1(b1)(b3)) is designed to summarize different combinations of voice modulation skills and provide speakers with hints for further improvements.

We propose a stacked bar chart based design to visualize information of voice modulation skills in a coarse-to-fine manner (Figure 1(b1)(b3)). The top part (Figure 1(b1)) presents general summary of retrieval results with respect to three conditions (i.e., not aligned, no technique and with technique). Each condition is encoded by a color or texture. The height of each segment of the stacked bar indicates the frequency of each type of the three conditions. For example, a tall dark gray bar implies that its corresponding word is popular for modulation, while a tall dashed-outlined bar indicates that the recommendation results of the word have insufficient support. The n-gram-based hierarchical visualization (Figure 1(b3)) summarizes the varied-length combinations of voice modulation skills in the retrieval results. The first row of stacked bars in Figure 1(b3) visualizes the voice modulation skills of each word, where a stacked bar chart is displayed under each word. The second row of the stacked bars in Figure 1(b3) shows the frequent combination of voice modulation skills for two adjacent words. The stacked bars are horizontally aligned at the center of the corresponding two words. The bars within each stacked bar are sorted in a descending order by the frequency of the corresponding voice modulation skill or the combination of voice modulation skills. When the user hovers over a voice modulation combination, its corresponding words will be highlighted in bold red.

The *voice technique table* at the bottom (Figure 3) shows a list of voice modulation skill sequences employed by the speaking examples in the TED benchmark dataset. These voice modulation skill sequences are sorted by their similarity with the voice modulation skill sequence extracted from the speaker's voice input. We use Hamming distance to measure the similarity between two sequences. The speaker can further explore interesting combinations of voice modulation skills by filtering techniques at the header of the table.

To ease comparison between the voice modulation skills employed by a speaker and the modulation skills used in the TED benchmark dataset, we come up with three designs to enhance the recommendation view. First, the modulation skills of a speaker (Figure 1(b2)), which are set as the baseline for comparison, are encoded by colored glyphs. Second, arrow markers are added in the n-gram based visualization to highlight the modulation skills used by both a speaker and the TED talks. Third, some buttons (👁 and ⊘) in Figure 1(b3) are added to help a speaker interactively set whether the n-gram-based visualization is shown or not. When it is hidden, the voice technique table will automatically move up and be positioned close to the glyphs of the voice modulation sequence of the speaker, enabling sequential comparative analysis of the voice modulation patterns.

[4] https://github.com/spotify/annoy
[5] https://www.spotify.com/



										Faster	Stress
	■	» s	s	■	■		■	» s	■	»	s
	■	s	■	■	«	»	s		«	»	s
	↓	»	■	»	■	» ↑	■		■	»	↑ s
	■	»	« s	»	«	»	■	»	«	» ↓	« s
	■	» ↑	■	↓	»	»	«	»	»	» s	s
	»	»	«	»	■	»	s	»	■	»	s
	■	■	■	■	■	»	■	» ↓	■	» ↓	s
	»	» s	■			» ↑	■			» ↑	» ↑ s
«		🟦	■	»		»	■	■		» ↑	« s
	» ↑	»	■	»	↓	■		s	↓	»	» s

Figure 3: The voice technique table. Users can filter sentences with specific voice modulations (e.g., faster and stress), then select some corresponding sentences for further exploration in the voice technique view by clicking the detail button.

Voice Technique View

When a user clicks on the voice modulation of his/her interest in the recommendation view, *VoiceCoach* can retrieve the TED talk segments that use the desired voice modulation skills and list them in the voice technique view. These TED talk segment examples are ranked by the sentence similarity of both sentence structural and semantic meanings between the user input and the TED talk segment examples. The retrieved TED talk segment examples can be highlighted in one-line mode (Figure 1(c1)) or multi-line mode (Figure 1(c2)), which enables the user to quickly locate and compare the local context of different voice modulation skills. When the user clicks on a word or a sentence ID in the voice technique view, the corresponding original TED talk voice will be played to give users a concrete understanding of the voice modulation skills.

Practice View

The practice view consists of three components: (1) a reference example showing the sentence with highlighted techniques that the user wants to practice (Figure 1(d1)), (2) a real-time feedback chart providing immediate quantitative feedback on the voice modulation skills employed by the speaker in his/her practice (Figure 1(d2)), and (3) a practice collection (Figure 1(d3)) storing and displaying all recorded practices To promote deliberate practice, the speaker is allowed to customize the words to focus on and techniques to be improved (Figure 1(d4)). To provide real-time and quantitative feedback on voice modulation, the feedback chart updates the real-time value of pitch (red solid line) and volume (dark blue area) of the current practice simultaneously, while the volume (light blue area) and pitch (green dashed line) of the previous practice are set as the baseline. Other vocal properties can also be inferred from the chart. For example, segments with zero volume indicate the pause and the speed of the volume wave suggests the speech rate.

Usage Scenario

We describe how Andy, an undergraduate student, utilizes *VoiceCoach* to practice and improve his voice modulation skills. Andy is preparing for a speech about negotiation skills, and he decides to take Isaac Newton's famous quote - "*Tact is the art of making a point without making an enemy*" - as a highlight of his talk. Therefore, he refers to *VoiceCoach* to perform deliberate practices on this quote.

After recording the script, he examines the recommendation view which shows the voice modulation skills he applied, in comparison with the recommended results. As shown in Figure 1(b2), he quickly notices that the voice modulation skills he used for several words (i.e., the color rectangles indicated by a black arrow on their left) are consistent with those recommendation results (e.g., *"tact", "art", "of" and "point"*), but there are also words where he does not use any voice modulation skills (indicated by the gray rectangles with a black black arrow) while TED speakers employed certain voice modulation skills. More specifically, an obvious exception of his speaking lays in the phrase "*making an enemy*", which is a key part of this quote but no voice modulation skills are adopted by Andy. He first tries to improve his speaking for "*an enemy*" by applying some voice modulation skills to them. Since the most frequent combination (i.e., » "faster" and ■ "no tech") does not apply any technique to the word *"enemy"*, he chooses the second most popular voice modulation combination, i.e., » "faster" and 🟪 "stress".

He decides to find an example with those techniques to mimic the voice modulation. He applies filters in the *voice technique table* to query sentences from the database. After he clicks the first returned result that has the highest similarity score with his phrase input, and listen to the example to develop a concrete idea about how a voice modulation combination of » "faster" and 🟪 "stress" should be, as shown in Figure 1(c).

Andy further uses the practice view to improve his speaking. As shown in Figure 4(a-c), his volume (the dark blue area) and pitch (the red line) are detected and shown in real time in each of his speaking practice for this quote. The corresponding volume (the light blue area) and pitch (the dotted green line) of his original speaking are used as a reference to show his improvement in each practice. The inconsistency of voice modulation for the phrase "*an enemy*" between each practice and the selected voice modulation combination is also highlighted in red dashed rectangles on the original text. From Figure 4(a-c), it is clear to see that Andy correctly applied the voice modulation combination of » "faster" and 🟪 "stress" into the phrase "*an enemy*" after three rounds of practice.

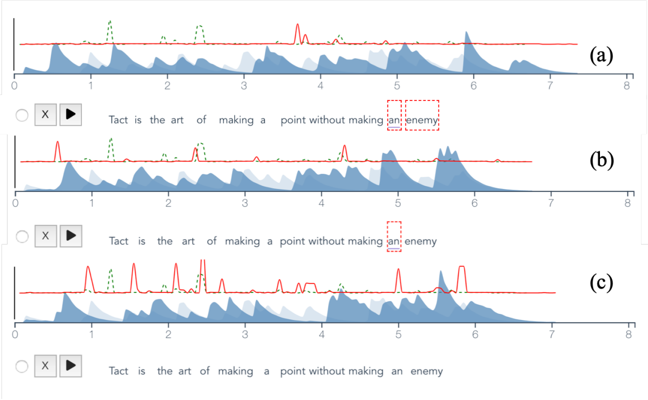


Figure 4: An illustration of multiple practices by a user, where the user focuses on practicing the phrase "an enemy". The decreasing number of the dashed red rectangles from (a) to (c) show the user's improvement of the voice modulation skills.

EXPERT INTERVIEW

We performed in-depth interviews with three domain experts (i.e., *E1-E3*), who also participated in our requirement analysis interviews, to evaluate the effectiveness and usability of *VoiceCoach*. We started the interviews by explaining the functions and visual encodings of *VoiceCoach*. A usage scenario was also introduced to showcase the usage of *VoiceCoach*. Then, we asked the experts to freely explore our system in a think-aloud manner and finish their exploration tasks, e.g., examine the recommendation results according to their voice input, select one desired modulation for further practice, and iteratively practice with on-the-fly quantitative feedback. After that, we collected their feedback on *VoiceCoach*. Each interview took about 1 hour, and all the interviews were recorded with the experts' consent. Overall, the experts showed great interest in *VoiceCoach*. Their feedback was summarized as follows.

Usefulness All the experts agreed that the evidence-based training in *VoiceCoach* could be helpful for novice speakers to improve their voice modulation skills. *E1* and *E3* mentioned that novice speakers are often not sure about what voice modulation skills to use and how to combine them in a new script, even though they may also already be aware of some high-level tips for voice modulation skills. They thought our recommendation strategy was new and clever. The voice modulation examples recommended by *VoiceCoach* provide speakers with evidence-based guidance. They can select suitable modulation skills for different sentences. *E2* pointed out that the on-the-fly feedback provided by *VoiceCoach* was more useful than that of traditional training, as there is usually just one coach with multiple students in a class of a traditional training program, making it difficult for the coach to provide sufficient and timely feedback to every student. *E1* commented that the quantitative feedback in *VoiceCoach* is very helpful for enabling a user to master voice modulation skills, as it provides the user with concrete real-time evaluations of their voice modulation skills during their practices. During the interviews, one interesting finding was that different coaches could have very different preferences for voice modulation. For instance, *E1* mentioned that pause is one of the most important and difficult voice modulation skills, thus his training often focused on pauses. On the contrary, *E2* confidently emphasized that the art of a successful speech lay in the good modulation of the volume and speed. Such observations further confirmed the importance of the example recommendations in *VoiceCoach*, which provide students with the flexibility to choose and follow the suitable "good" voice modulation examples.

Visual designs and usability All three experts appreciated the evidence-based training provided by *VoiceCoach*. They confirmed that the overall visualization designs were intuitive and easy to understand. For the recommendation view, *E1* said that it demonstrated the diversity of voice modulation skills. *E2* pointed out *"Though the recommendation view seems to be the most complex view of VoiceCoach at first glance, I can quickly understand and learn how to use it after your brief introduction.*" All experts mentioned that most of the top-ranked recommendation examples in the voice technique view made sense to them. By clicking the corresponding sentence in the voice technique view, they could conveniently check how those voice modulation skills were used by the TED speakers. For the practice view, they agreed that the real-time feedback charts, as well as the highlighted text boxes, help them recognize the difference between different practices. In addition, the experts were highly impressed by the convenient and smooth interactions of *VoiceCoach*.

Limitations and suggestions Despite the overall positive feedback from the experts, they also pointed out some limitations of *VoiceCoach* and gave us insightful suggestions on it. *E2* said that *VoiceCoach* currently only recommended "good" voice modulation examples for speakers to follow, while speakers could also benefit from negative examples. By informing them of "bad" modulation such as a monotone voice, they could easily know what mistakes they should avoid. *E2* suggested that it would be interesting to classify the speakers into different types (e.g., fast speaker vs. slow speaker, soft speaker vs. loud speaker) and to deliberately recommend voice modulation examples to them (e.g., recommend fast speaking examples to slow speakers and loud speaking examples to soft speakers). Due to the limited voice modulation datasets that are available, we have left this as part of our future work.

USER STUDY

We conducted a well-structured user study to evaluate the effectiveness and usability of *VoiceCoach* for the training of voice modulation skills. Since the concrete voice modulation examples (the recommendation view and voice technique view) and the immediate and concrete feedback (the practice view) are the two major desirable functions of *VoiceCoach*, we designed the user study with an emphasis on these two aspects. Specifically, we aimed to answer the following questions:

- **Recommendation helpfulness:** How helpful is our system in finding appropriate voice modulation examples to guide the practice?
- **Effectiveness of immediate feedback:** How effective is our system for improving participants' voice modulation skills in terms of getting quick and quantitative feedback?
- **Overall usability and effectiveness:** Is *VoiceCoach* effective for improving participants' skills of voice modulation and is it easy to use?

Participants

We recruited 18 university students (4 females, $age_{Mean} = 23$, $age_{SD} = 2.52$) from a local university through word-of-mouth and flyers. They came from different backgrounds, including chemistry, math, computer science, mechanical engineering, and finance. Each participant received $17. All the participants had experiences of public speaking, but none of them had attended any professional training of voice modulation. They have all expressed the desire to improve their presentations and an eagerness to improve their skills of voice modulation. All the participants had normal vision and hearing.

Experiment Design

Before the study, we worked together with the coaches (*E1*, *E2*) and selected 13 sentences (*S1-S13*) as training examples. These examples have been popularly used in their training programs. Our user study consisted of four sessions.

In the first session, we introduced the purpose and the procedures of our study. After that, we illustrated the skills of voice modulation that we have mentioned in this paper with example videos and gave them some general tips about the usage of such skills. After they had grasped the concepts of voice modulation, we demonstrated how to use our system.

In session two, we asked participants to freely explore *VoiceCoach* in a think-aloud manner with four sample sentences *(S1-S4)*, which aimed to familiarize them with the system.

In session three, participants were presented with another five sentences *(S5-S9)* and asked to examine the results generated by the recommendation view and the voice technique view. The tasks were to explore the recommended voice modulation skills and their corresponding words or phrases. Meanwhile, they were requested to report how many recommended examples they believed were relevant in terms of sentence structure and voice modulation skills among the top five retrieval results in the voice technique view. Their click activities were also captured for further analysis. At the end of session three, participants needed to complete a questionnaire consisting of 11 questions, where they evaluated the recommendation results (*Q1-11*) in a 7-point Likert scale, as shown in Table 3.

In Session four, we compared our system with a baseline system using another set of four sentences *(S10-S13)*, where the baseline system was a simplified version of *VoiceCoach* by removing the feedback generated from the practice view and only reserved the functions of recording and playback. This simplified system only allowed participants to listen to his/her own audios and make the adjustment accordingly, which simulated real-world practice. For each sentence, participants were asked to practice it with the same pre-defined instructions using either *VoiceCoach* or the baseline system. To minimize the learning effect, we evaluated the two systems in a counterbalanced order. Also, a questionnaire of 5 questions (*Q12-Q16*) in Table 3 with a 7-point Likert scale was to be finished afterwards. After four sessions, we conducted a post-study survey with the participants, during which we had them finish (*Q17-Q25*) in Table 3 and answer some general questions about their experience of the training. The whole study lasted about 90 minutes.

Results and Analysis

Evaluation on Recommendation Results

We analyzed the user-generated data (i.e., click data, report of the number of relevant examples in the top 5 retrieved results) and the ratings from the questionnaire (*Q1-Q11* in Table3). The results show the usability and effectiveness of the recommendation view and the voice technique view. On average, participants clicked on modulation combinations in the recommendation view about 4.27 times ($SD = 1.27$) before they settled down to a desirable combination, and 4.21 ($SD = 1.21$) of top 5 retrieved results displayed in the voice technique view satisfied the participants' needs. The relevance rate of the recommended examples was 89%. Besides, most participants showed positive responses to the recommendation results, especially in terms of decision making and usability. Interestingly, one participant (5.6%) disagreed about the relevance of the recommendation results and one participant (5.6%) was neutral about the intuitiveness of the visual design of recommendation view. The summary of the feedback is shown in Figure 5. In addition, we had our experts *E1, E2* go through all the chosen techniques of participants, and they found them reasonable and applicable.

Table 3: Three questionnaires designed for Sessions three, four and the post-study survey. Assessment of the quality of of the recommendation results in four aspects: informativeness (*Q1-Q3*), visual design (*Q4-Q6*), decision making (*Q7-Q9*), usability (*Q10-Q11*). Assessment of the effectiveness of vocal practice: self-awareness (*Q12-Q13*), self-adjustment (*Q14-Q15*), self-reported evaluation (*Q16*). Participants' feedback about *VoiceCoach*: voice modulation (*Q17-Q18*), system components (*Q19-Q22*), usability (*Q23-Q25*).

Q1	The information needed is easy to access.
Q2	The information of other speakers' voice modulation is rich.
Q3	The results are relevant in terms of structural information and selected techniques.
Q4	The visual design is intuitive.
Q5	The visual design is helpful to identify the differences.
Q6	The visual design is helpful to find specific combinations.
Q7	The visual design is helpful to find what needs to be improved.
Q8	I am confident that I find suitable techniques to practice.
Q9	The views help me understand the usage of voice modulation skills.
Q10	It was easy to learn.
Q11	It was easy to use.
Q12	Listening to my own audio is helpful for self-awareness of my performance.
Q13	Technique labels are helpful for self-awareness of my performance.
Q14	Listening to my own audio is easy for adjustment of my voice.
Q15	Real-time quantitative feedback is easy for adjustment of my voice.
Q16	I am more satisfied with VoiceCoach than simply listening to audio.
Q17	The system helps me know where my speaking voice can be improved.
Q18	The system helps me gain diverse information about voice modulation.
Q19-Q22	Technique labels, visual summary, real-time feedback, parallel alignment are helpful.
Q23	It was easy to learn.
Q24	It was easy to use.
Q25	I would recommend this system to others.

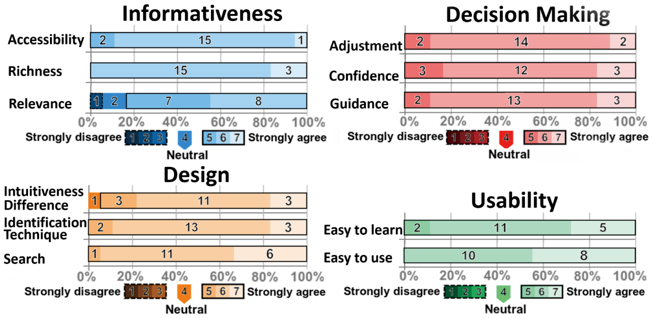


Figure 5: The results of questionnaire about helpfulness of recommendation in four aspects, including informativeness (*Q1-Q3*), decision making (*Q7-Q9*), design (*Q4-Q6*), and usability (*Q10-Q11*).

Evaluation on Practice View

We ran Wilcoxon signed-rank tests on the feedback for self-awareness and self-adjustment during the voice modulation practice in Session four, to compare the effectiveness between *VoiceCoach* and the baseline. The result (Figure 6 (a)) shows a significant difference in the self-awareness scores ($p < 0.001$, $Z = -3.75$), which indicated that *VoiceCoach* better helped participants understand their vocal performance ($Mean = 6.00$, $SD = 0.77$) compared with listening to personal audio records ($Mean = 3.17$, $SD = 1.47$). Significant differences in the self-adjustment scores were also observed ($p < 0.001$, $Z = -3.70$), showing that *VoiceCoach* better assisted participants in adjusting the participants' voice ($Mean = 5.89$, $SD = 0.58$) compared with the baseline method ($Mean = 2.72$, $SD = 1.27$).

Furthermore, the ratings of the questionnaire (Figure 6 (b)) suggest that all participants prefer *VoiceCoach* for practicing.

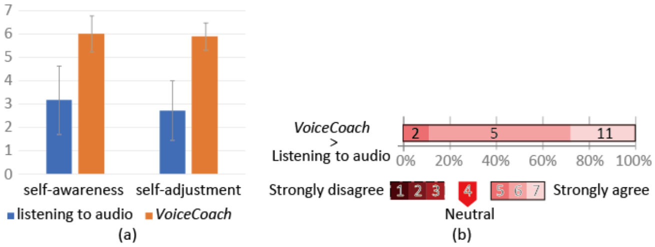


Figure 6: The results of questionnaire about user experience of practice. (a) Comparison of self-awareness and self-adjustment between *VoiceCoach* and the baseline. (b) Participants' responses to *Q16*.

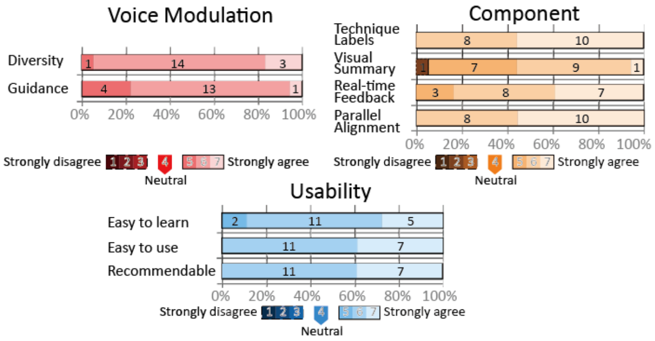


Figure 7: The results of participants' feedback about system in terms of voice modulation (*Q17-Q18*), component design (*Q19-Q22*) and usability *(Q23-Q25)*.

Overall feedback on VoiceCoach

We collected feedback from the questionnaire (*Q17-Q25* in Table 3) in the post-study survey. The result (Figure 7) shows that all participants found that *VoiceCoach* enriched their knowledge about voice modulation and helped them improve vocal skills. Also, most participants agreed that the four core components of the system were useful, especially technique labels and parallel alignment. One participant (*P2*, Female, 24) found the visual summary less helpful, because it took her a while to understand the design of recommendation view. In general, participants claimed that *VoiceCoach* had good usability.

During the post-study interview, we asked participants about their experience of using *VoiceCoach* and the new knowledge of voice modulation they had learnt. We also collected comments and suggestions for the user panel, the recommendation view, the practice view and the voice technique view.

Training experience In general, all participants felt excited about *VoiceCoach*. One participant (*P8*, Male, 20) strongly believed *"(VoiceCoach) would be a helpful training tool for the speakers to practice their voice anytime anywhere."* Five participants mentioned that by observing the differences between their voice and experts' voice, they gained insights about their inadequate usage of voice modulation. One of them (*P4*, Male, 23) commented *"...comparing with TED talkers made me realize that I normally spoke very fast and did not vary the speed much."* Another participant (*P13*, Female, 18) was amazed by the power of a pause after her training: *"I can't believe that a pause has such magic to make my voice sound so dramatic."*

Visual designs The visual design of the system seemed intuitive to most participants, especially the technique labels for feedback. One (*P18*, Male, 22) said *"The technique labels were simple and compact. Instead of listening to audio myself, I could quickly discover the sequential patterns (of voice modulation) in audio by these labels."* Many participants found the arrow markers in the recommendation view helpful for identifying the differences between their voice and others' in all levels of voice modulation combinations. Interestingly, we noticed that some of them held contradictory opinions towards the sentence-level summary of voice techniques. One participant (*P2*, Female, 24) found it less useful than n-gram-based visualization: *" The sentence I selected in the voice technique table did not seem relevant to my sentence."* While another (*P13*, Female, 18) thought *" The examples recommended in the table were so helpful for learning."* The conflicts may be caused by the limited size of our dataset. One participant (*P13*, Female, 18) described the practice as a voice game: *"It was very interesting to see the real-time feedback of my voice on the screen. It reminded me of where and when I should make adjustments."* Also, she expressed her difficulty in focusing on several dimensions simultaneously during the practice.

Interactions Overall, participants enjoyed the rich and effective interactions which helped them explore recommendation results. Many participants mentioned about the convenience of parallel alignment and the auto-focus of the corresponding contexts of selected techniques in the voice technique view. One (*P14*, Male, 28) commented *"The voice technique view saved my time. I could discover combinations of interests by one glance at the table."* Another one (*P4*, Male, 23) added *"It was very considerate of you to let me listen to the words I want with a simple click. I didn't bother to listen to the whole sentence."* Many participants agreed that it was beneficial to let them focus on specific words and modify the unwanted techniques, which eased the whole process of practices. After the experiment, two participants showed their strong interests in *VoiceCoach* and spent extra time on exploring our system.

Evaluation by coaches

To further determine the training effectiveness of *VoiceCoach*, we invited the two aforementioned coaches (*E1, E2*) to evaluate the speakers' performance. Specifically, we recruited another 24 university students (9 females, $age_{Mean} = 24$,

$age_{SD} = 2.47$) from our university and randomly divided them into two groups (*G1*, *G2*). Participants were asked to practice their voice modulation skills based on the same script with or without *VoiceCoach*. The script was a 30-second speech opening pre-selected by *E1* and *E2*, and *G1* was set as the control group. After that, coaches evaluated speakers' final audio presentation in terms of diversity, coherence, and expressiveness of voice modulation with a 7-point Likert scale. Both coaches were blind to the study condition.

We analyzed the performance scores of *G1* and *G2* using Wilcoxon signed-rank tests. There was a significant difference ($p = 0.03, Z = -2.15$) between *G1* ($Mean = 4.17, SD = 1.03$) and *G2* ($Mean = 5.08, SD = 1.08$), which indicates that *VoiceCoach* better helped improve voice modulation skills.

DISCUSSIONS AND LIMITATIONS

VoiceCoach is designed to provide novice speakers with evidence-based training of voice modulation skills. Our in-depth expert interviews and user study provide support for the usefulness, effectiveness, and usability of *VoiceCoach* in facilitating the training of voice modulation skills. However, there are still several key aspects that need further discussions.

Lessons learned We summarize the important lessons learned from our system implementation, and evaluation studies. 1) *Design a progressive learning process for skill acquisition.* During our design process, experts pointed out that it is challenging for novice speakers to apply all kinds of voice modulation skills to one sentence and to master new skills in one try. To ease the training process and to improve learning efficiency, our system promotes deliberate and iterative voice modulation practice on words of interests. During our user study, participants acknowledged the design of practice view as helping them focus on specific parts of the sentence and gradually improving their skills by highlighting issues in their previous practices. Thus, we expect that the system should develop strategies of breaking down the overall training goal into small tasks and giving users step-by-step instructions on the tasks. 2) *Turn practice into a game.* During the user study, we observed that several participants spent extra time interacting with voice curves in the practice view. They tried all the example sentences with different recommended modulation skills, and reported that interacting with real-time feedback was like playing a game. This indicates that adding interesting designs in the training system can increase user engagement, benefiting successful learning of skills. 3) *Provide flexible and personalized training.* In the expert interviews, we found that coaches had very different preferences for voice modulation skills, which may lead to a biased training of specific types of voice modulation in the traditional methods of public speaking training, and a failure to meet the needs of speakers from different backgrounds. These subjective biases provide support for the importance of a flexible and personalized training.

Effectiveness and usability evaluations Our current evaluations consist of in-depth interviews with domain experts and user studies with university students, which can provide support for the evaluation of the effectiveness and usability of *VoiceCoach*. The current system will be deployed to the public speaking training platforms of our industry collaborator. With more participants from diverse background, it will further evaluate and verify the effectiveness and usability of *VoiceCoach*.

Technical limitations of *VoiceCoach* First, we use TED talks as the benchmark dataset. Though 2,623 high-quality speeches are included, they may still do not cover all the "good" speeches in different domains. For example, the desirable voice modulation skills for an academic talk can be different from that for a business talk, but there are not many academic talks in the TED dataset. Second, *VoiceCoach* currently focuses on recommending "good" voice modulation examples and how to help speakers to learn from "bad" examples is not explored, which, as mentioned by the coaches in our expert interviews, may be also beneficial to the voice modulation training. Third, our current recommendation of voice modulation examples is mainly based on the similarity of sentence structures, which does not consider the preferences of different users in different speaking scenarios.

CONCLUSION AND FUTURE WORK

In this paper, we present *VoiceCoach*, an interactive evidence-based training system for voice modulation skills in public speaking. By working closely with professional communication coaches from our industry collaboration company in the past eight months, we have identified two of the most important major requirements of effective voice modulation training: *concrete and personalized guidelines* and *on-the-fly* feedback. Accordingly, we analyzed 2,623 high-quality TED speeches and recommend voice modulation examples to users based on the sentence structure similarity between the voice input and the TED speech segments, providing users with evidence-based hints on improvements of their vocal skills. *VoiceCoach* further enables quantitative and immediate feedback, through comparing the volume, pitch, and speed of users' voice input with their prior practice, to guide their further improvement on voice modulation skills. Our semi-structured expert interviews and user study with university students provide support for the good usability and effectiveness of *VoiceCoach* in helping novice speakers with the training of voice modulation skills.

In future work, we would like to extend the current benchmark dataset by including the speeches in different domains (e.g., academic talks, public campaigns), and further improve the applicability of *VoiceCoach*. It would also be interesting to collect "bad" examples of voice modulation and improve the current system by showing negative examples as well to users, informing them of the voice modulation mistakes they should avoid. Furthermore, we plan to invite more participants with more diverse backgrounds, to further validate the usability and effectiveness of *VoiceCoach* in helping novice speakers with evidence-based training of voice modulation skills.

ACKNOWLEDGEMENTS

The authors would like to thank the experts and participants for their help in the project, as well as the anonymous reviewers for their valuable comments. This project is funded by a grant from ITF UICP (Project No. UIT/142).

REFERENCES

[1] Tony Bergstrom and Karrie Karahalios. 2007. Conversation Clock: Visualizing audio patterns in co-located groups. In *Proceedings of the Hawaii International Conference on System Sciences*. IEEE, 78–78.

[2] Jordan L Boyd-Graber, Sonya S Nikolova, Karyn A Moffatt, Kenrick C Kin, Joshua Y Lee, Lester W Mackey, Marilyn M Tremaine, and Maria M Klawe. 2006. Participatory design with proxies: developing a desktop-PDA system to support people with aphasia. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 151–160.

[3] Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, and others. 2018. Universal sentence encoder. *arXiv preprint arXiv:1803.11175* (2018).

[4] Ionut Damian, Chiew Seng Sean Tan, Tobias Baur, Johannes Schöning, Kris Luyten, and Elisabeth André. 2015. Augmenting social interactions: Realtime behavioural feedback using social signal processing techniques. In *Proceedings of the International Conference on Human Factors in Computing Systems*. ACM, 565–574.

[5] Fiona Dermody and Alistair Sutherland. 2016. Multimodal system for public speaking with real time feedback: a positive computing perspective. In *Proceedings of the International Conference on Multimodal Interaction*. ACM, 408–409.

[6] Joseph A DeVito. 2003. *The essential elements of public speaking*. Allyn and Bacon.

[7] Jonathan Foote. 1999. Visualizing music and audio using self-similarity. In *Proceedings of the International Conference on Multimedia*. ACM, 77–80.

[8] C. Gallo. 2014. *Talk Like TED: The 9 Public Speaking Secrets of the World's Top Minds*. Pan Macmillan. `https://books.google.com.hk/books?id=K3v8AgAAQBAJ`

[9] Jiawei Han, Jian Pei, and Yiwen Yin. 2000. Mining frequent patterns without candidate generation. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, Vol. 29. ACM, 1–12.

[10] Julia Bell Hirschberg and Andrew Rosenberg. 2005. Acoustic/prosodic and lexical correlates of charismatic speech. In *Proceedings of European Conference on Speech Communication and Technology*. Lisbon, 539–546.

[11] Mohammed Ehsan Hoque, Matthieu Courgeon, Jean-Claude Martin, Bilge Mutlu, and Rosalind W Picard. 2013. Mach: My automated conversation coach. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing*. ACM, 697–706.

[12] Toastmasters International. 2011. Your speaking voice. `https://www.toastmasters.org/Resources/Your-Speaking-Voice`. (2011). Last accessed on 2019-09-20.

[13] Dasaem Jeong and Juhan Nam. 2016. 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, 25–32.

[14] Kazutaka Kurihara, Masataka Goto, Jun Ogata, Yosuke Matsusaka, and Takeo Igarashi. 2007. Presentation sensei: a presentation training system using speech and image processing. In *Proceedings of the International Conference on Multimodal Interfaces*. ACM, 358–365.

[15] Stephen Lucas and Paul Stob. 2004. *The art of public speaking*. McGraw-Hill New York.

[16] Piet Mertens. 2004. The prosogram: Semi-automatic transcription of prosody based on a tonal perception model. In *Proceedings of the International Conference on Speech Prosody*. Nara, Japan, 23–26.

[17] Leslie Milton and Christine Lu. 2015. VerseVis: Visualization of spoken features in poetry. *University of Maryland, Tech. Rep* (2015), 1–9.

[18] Fabian Mörchen, Alfred Ultsch, Mario Nöcker, and Christian Stamm. 2005. Databionic visualization of music collections according to perceptual distance. In *Proceedings of the International Society for Music Information Retrieval*. 396–403.

[19] Chris Muelder, Thomas Provan, and Kwan-Liu Ma. 2010. Content based graph visualization of audio data for music library navigation. In *Proceedings of the International Symposium on Multimedia*. IEEE, 129–136.

[20] Arina Nikitina. 2011. *Successful public speaking*. Bookboon.

[21] Alp Öktem, Mireia Farrús, and Leo Wanner. 2017. Prosograph: a tool for prosody visualisation of large speech corpora. In *Proceedings of the Annual Conference of the International Speech Communication Association*. 809–810.

[22] Elias Pampalk, Andreas Rauber, and Dieter Merkl. 2002. Content-based organization and visualization of music archives. In *Proceedings of the ACM International Conference on Multimedia*. ACM, 570–579.

[23] Rupal Patel and William Furr. 2011. ReadN'Karaoke: visualizing prosody in children's books for expressive oral reading. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 3203–3206.

[24] Katarzyna Pisanski, Valentina Cartei, Carolyn McGettigan, Jordan Raine, and David Reby. 2016. Voice modulation: a window into the origins of human vocal control? *Trends in Cognitive Sciences* 20, 4 (2016), 304–318.

[25] Steve Rubin, Floraine Berthouzoz, Gautham J Mysore, and Maneesh Agrawala. 2015. Capture-time feedback for recording scripted narration. In *Proceedings of the Annual ACM Symposium on User Interface Software & Technology*. ACM, 191–199.

[26] RA. Schmidt, DE. Young, S. Swinnen, and DC. Shapiro. 1989. Summary knowledge of results for skill acquisition: support for the guidance hypothesis. *Journal of Experimental Psychology: Learning, Memory and Cognition* 15, 2 (1989), 352–359.

[27] Jan Schneider, Dirk Börner, Peter Van Rosmalen, and Marcus Specht. 2015. Presentation trainer, your public speaking multimodal coach. In *Proceedings of the International Conference on Multimodal Interaction*. ACM, 539–546.

[28] Jan Schneider, Dirk Börner, Peter Van Rosmalen, and Marcus Specht. 2017. Presentation Trainer: what experts and computers can tell about your nonverbal communication. *Computer Assisted Learning* 33, 2 (2017), 164–177.

[29] Prem Seetharaman, Gautham Mysore, Bryan Pardo, Paris Smaragdis, and Celso Gomes. 2019. VoiceAssist: Guiding Users to High-Quality Voice Recordings. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*. ACM, 1–6.

[30] Eva Strangert. 2005. Prosody in public speech: analyses of a news announcement and a political interview. In *Proceedings of the European Conference on Speech Communication and Technology*. 3401–3404.

[31] Eva Strangert and Joakim Gustafson. 2008. What makes a good speaker? subject ratings, acoustic measurements and perceptual evaluations. In *Proceedings of the Annual Conference of the International Speech Communication Association*. KTH, Speech, Music and Hearing, TMH, 1688–1691.

[32] Hiroki Tanaka, Sakriani Sakti, Graham Neubig, Tomoki Toda, Hideki Negoro, Hidemi Iwasaka, and Satoshi Nakamura. 2015. Automated social skills trainer. In *Proceedings of the International Conference on Intelligent User Interfaces*. ACM, 17–27.

[33] M Iftekhar Tanveer, Emy Lin, and Mohammed Ehsan Hoque. 2015. Rhema: A real-time in-situ intelligent interface to help people with public speaking. In *Proceedings of the International Conference on Intelligent User Interfaces*. ACM, 286–295.

[34] Ha Trinh, Reza Asadi, Darren Edge, and T Bickmore. 2017. RoboCOP: A Robotic Coach for Oral Presentations. In *Proceedings of the ACM International Conference on Interactive, Mobile, Wearable and Ubiquitous Technologies*, Vol. 1. ACM, 1–24.

[35] TJ Tsai. 2015. Are You TED Talk material? Comparing prosody in professors and TED speakers. In *Proceedings of the Annual Conference of the International Speech Communication Association*. ISCA, 2534–2538.

[36] Haipeng Zeng, Xingbo Wang, Aoyu Wu, Yong Wang, Quan Li, Alex Endert, and Huamin Qu. 2019. EmoCo: Visual Analysis of Emotion Coherence in Presentation Videos. *IEEE Transactions on Visualization and Computer Graphics* 26, 1 (2019), 927–937.

[37] Ru Zhao, Vivian Li, Hugo Barbosa, Gourab Ghoshal, and Mohammed Ehsan Hoque. 2017. Semi-automated & collaborative online training module for improving communication skills. In *Proceedings of the ACM International Conference on Interactive, Mobile, Wearable and Ubiquitous Technologies*, Vol. 1. ACM, 1–20.