systems

MDPI

Article

Effective Presentation Speech Support System for Representing Emphasis-Intention

Tomoko Kojiri * and Takaya Kaji

Received: 29 September 2015; Accepted: 15 December 2015; Published: 23 December 2015
Academic Editors: Nguyen-Thinh Le, Jon Mason, Christian Gütl and Kiyoshi Nakabayashi

Faculty of Engineering Science, Kansai University, 3-3-35, Yamate-cho, Suita, Osaka 564-8680, Japan; k443110@kansai-u.ac.jp
* Correspondence: kojiri@kansai-u.ac.jp; Tel.: +81-6-6368-0968

Abstract: A research presentation integrates slides and speech. If these two aspects do not represent the same intention, the presentation will probably fail to effectively explain the presenter's intention. This paper focuses on the representation of the critical contents in a presentation. In an effective speech, the speaker adds more intonation and stress to emphasize the importance of the slide contents. Audiences recognize that important contents are those that are explained in a stronger voice or that are said after a short pause. However, in ineffective speeches, such voice effects do not always correspond to the important contents that are indicated by slides. On slides, the important contents are represented by levels of text indentation and size, color, and animation. This research develops a presentation speech support system that estimates important contents from slides and voices that might be recognized by audiences and extracts numerical differences. In addition, the system provides comments and feedback to improve speeches.

Keywords: presentation speech support; emphasis-intention; recognition of speech intonation

1. Introduction

Scientists and researchers often explain their ideas and work using presentation support tools such as Microsoft PowerPoint. During presentations, in addition to slides that visually represent their contents, we speak about and explain them by providing verbal explanations as well as gestures and body language. The intonation and the speed of verbal explanations emphasize the important contents in the slides and provide a rhythm for the speech. However, since many presenters fail to effectively express themselves through their speech, including the importance of slide contents, they fail to successfully explain their work or ideas to audiences. In this research, emphasis-intention is our term for the motivation that emphasizes important contents. The objective of our research is to support presenters so that they can give speeches that represent the emphasis-intentions of important slide contents. Currently, we are focusing on research presentations in the computer science field, where slides contain the main ideas of topics and the oral speech adds complementary explanations. Our target presenters are those who have already prepared slides and want to practice presenting them by learning to add appropriate speech techniques for more adequate explanations.

Despite the critical importance of speech in presentations, much research that supports presentations focuses on slide creation [1–3]. To help presenters create additional explanations in speech, Kojiri *et al.* developed a system that analyzes and visualizes the relations among slides that audiences can recognize [4]. By monitoring the visualized relations, slides that need additional explanation are highlighted to represent their relations. This research supports the creation of explanation sentences in speech. Okamoto *et al.* developed a presentation rehearsal support system [5] with which both slides and speech can be improved if rehearsal audiences point out inappropriate situations or points. However, this system requires an audience. Presenters can also learn how to speak more effectively or powerfully from books or web pages [6–8]. However, they have trouble determining whether their speeches are actually following such effective presentation techniques.

We have developed a support system for making presentation speech more powerful and effective. In our system, presenter speeches are recorded, and emphasis-intentions are extracted by analyzing intonations. It also extracts the important slide contents from slide layout information and analyzes whether the speech effectively represents the slides' emphasis-intentions. Then it offers feedback for improving the speech, if necessary.

The final goal of our research is detecting the exact slide contents that are emphasized by speech. However, this is not easy because analyzing the target slide contents from speech requires sophisticated natural language processing technologies. As a first step, we focus on the numbers of emphasis-intentions in each slide that are extracted from the speech and the slides. If the numbers extracted from the slides and the speech are different, the speech failed to appropriately represent the emphasis-intentions, and the system offers negative feedback.

2. Approach of Presentation Speech Support System

2.1. Representation of Emphasis-Intention

The "Present In English" website targets presenters for whom English is a second language and argues that intonation and rhythm are critical to making audiences understand the importance of contents [9]. Presenters should not speak monotonously; they need to change their voice volume and speed, based on the contents. If their intonation is inappropriate or awkward, audiences will probably not understand the critical slides; volume and speed must reflect the importance of the slide contents.

Audiences can obviously hear the contents of slides more clearly if they are spoken in louder voices. Therefore, audiences will infer that contents spoken in a louder voice are more important. On the other hand, one slide's explanation consists of several sentences, and presenters usually do not provide all of the explanations at once. Since pauses of 1 or 2 s attract attention, the contents after a brief pause are emphasized and deemed important [10].

The importance of slide contents is shown by the attributes of the slide objects, which are the components in slides that give visual effects. The text objects on slides contain many contents, and inclusive relations between each piece of content are often represented by indentation levels. Since more important contents have higher indentation levels [11,12], text at the first indentation level is more important. In addition, text attachments such as color, italics, or underlining provide emphasis. Regardless of the indentation level, texts with such attachments are also regarded as more important. On the other hand, objects other than text are often animated to capture the audience's attention [13]; objects with animation are also emphasized.

2.2. Overview of Presentation Speech Support System

Well-intoned speech gives emphasis-intention to the slide contents. That is, when explaining a slide's contents with emphasis-intention, the presenter's voice becomes louder. A short pause can also be placed before explaining them. When explaining more mundane slide contents, the voice remains calmer or flatter with shorter or fewer pauses.

This research has developed a presentation speech support system that extracts emphasis-intentions and gives feedback if the numbers of emphasis-intentions extracted from the slides and the speech are different. Figure 1 shows an overview of the system. The solid lines express the control flow, and the dashed lines represent the data flow. A slide file and the sound of the presenter's speech are input. At the extraction mechanism of the emphasized slide contents, the slides are decomposed into objects and those with emphasis-intentions are extracted as emphasized slide contents. In the extraction mechanism of the emphasized speech parts, the speech sound is analyzed, and such emphasized parts are extracted as loud voices and longer pauses. Then the sound with the emphasized parts is stored as emphasized speech parts. In the extraction mechanism of the intention differences, the numbers of the emphasized slide contents and the emphasized speech parts are compared. If the numbers are not the same, feedback that suggests improving the speech is given to the presenter. It also visualizes slides with emphasized slide contents and speech waveforms with emphasized speech parts so that presenters can check if their intentions were extracted correctly.

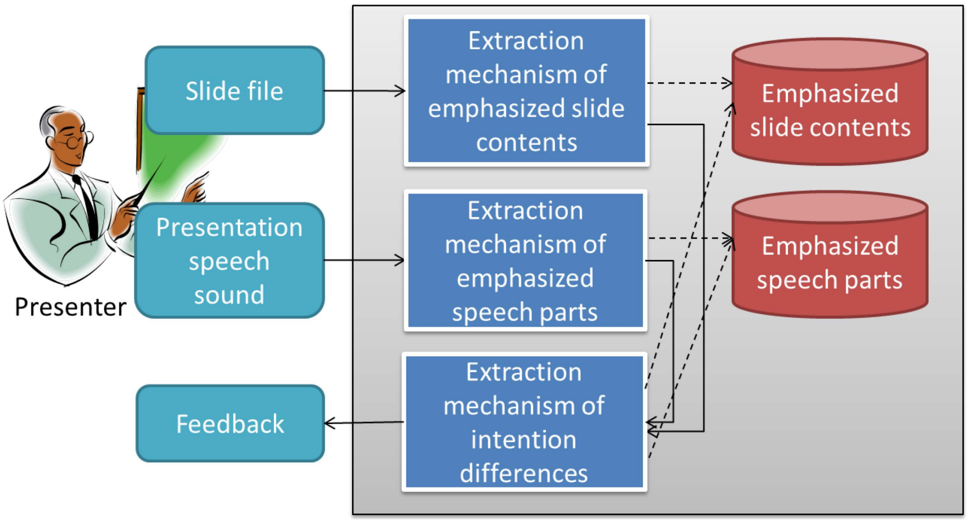

Slide file
Presenter
Presentation speech sound
Feedback
Extraction mechanism of emphasized slide contents
Extraction mechanism of emphasized speech parts
Extraction mechanism of intention differences
Emphasized slide contents
Emphasized speech parts

Figure 1. System overview.

3. System Mechanisms

3.1. Extraction Mechanism of Emphasized Slide Contents

We defined the following texts/objects as emphasized slide contents, which are determined by the attributes of the slide objects:

- texts with the highest indentation levels;
- texts with text attachments;
- objects to which animation is attached.

Our research focuses on the presentation tool of Microsoft PowerPoint. Here we explain how to detect the emphasized slide contents with C# programming language. If we handle the files in the PowerPoint format with C# language, all of the objects are handled as a Shape Interface in Microsoft.Office.Interop.PowerPoint. By checking the properties of the Animation Settings, the existence of animation in the objects is identified.

On the other hand, the text contents are handled as a Shape.TextFrame.TextRange class. Text at the highest indentation level can be acquired by indicating the IndentLevel property as 1, and text attachments are grasped by the Font property of its Text.

3.2. Extraction Mechanism of Emphasized Speech Parts

The emphasized speech parts are extracted from the speech waveform based on the following methods:

- **Speech in a loud voice** is a place where a presenter speaks with a loud voice for a certain time. Loudness is grasped by the waveform's amplitude. The speech parts whose amplitudes exceed a threshold for a certain time are regarded as a loud voice. Figure 2 shows an example of detecting a loud voice. The *x* axis represents the sampling number, and its sampling rate is 10,000.

First, waves whose amplitudes exceed threshold *a* are detected as loud waves. Second, if the next loud wave is found in time *b*, the waves might be part of the same sound and are regarded as one loud wave. If the length of the loud wave exceeds time *r*, the waves form a loud voice. Currently, thresholds *a*, *b*, and *r* are heuristically set to 0.6, 4000 (*i.e.*, 0.4 s), and 100,000 (*i.e.*, 10 s).

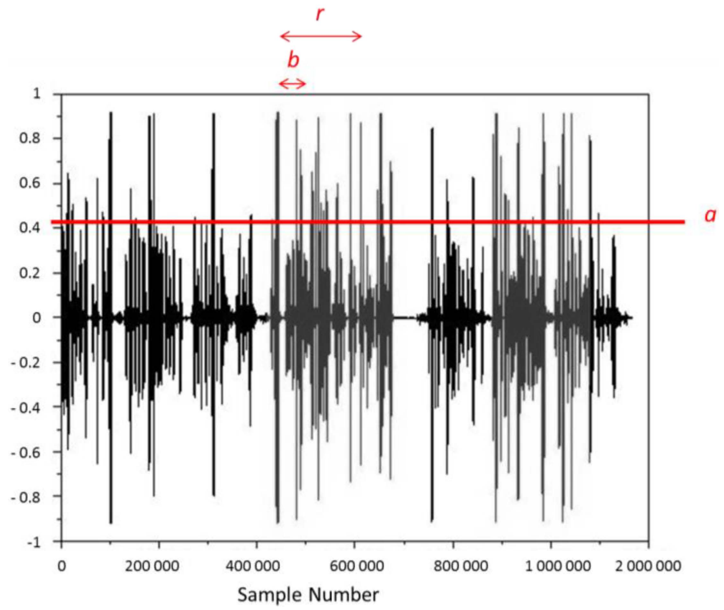


Figure 2. Parameters for detecting loud voices.

- **A pause** is a place where a presenter is not talking for a very brief moment. Thus, the speech parts whose amplitudes are below a threshold for a certain time are regarded as a pause. Figure 3 shows an example of detecting a pause. The *x* axis also represents the sampling number, and its sampling rate is 10,000. First, the waves whose amplitudes are below *x* are detected as silent waves. If silent waves continue for time *y*, they are defined as a pause. Currently, thresholds *x* and *y* are set to 0.05 and 15,000 (*i.e.*, 1.5 s).

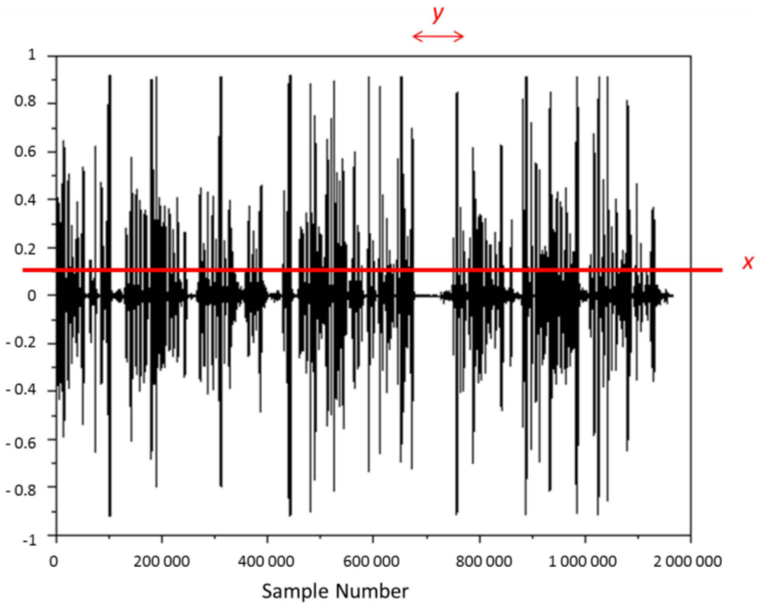


Figure 3. Parameters for detecting pauses.

3.3. Extraction Mechanism of Intention Differences

If the speech gives appropriate intonations to important slide contents, the numbers of the emphasized slide contents and emphasized speech parts will be the same. If the numbers are not the same, presenters will be encouraged to improve their speech techniques.

We extract the following three differences and generate feedback. Here, [slide] denotes the number of emphasized slide intentions, [loudness] indicates the number of louder speech parts, and [pause] shows the number of pauses in the speech.

- [slide] > [loudness]: Some important slide contents were *not* explained in a loud voice. The system generates feedback that suggests making the voice louder for important contents: *"When you are discussing important slide contents, speak more loudly."*
- [slide] < [loudness]: Some slide contents of lesser importance were explained in a loud voice. This unnecessarily loud voice might mislead the listeners into believing that relatively prosaic slide contents are important. Therefore, the system generates feedback that suggests avoiding unnecessarily loud voices: *"If you are NOT discussing important slide contents, don't add power to your voice."*
- [slide] > [pause]: Before some important slide contents, no pause or too-short pauses were made. The system provides feedback that suggests making more (or longer) pauses to draw more attention to subsequent explanations: *"If you are talking about important slide contents, make a slightly longer pause before saying them."*

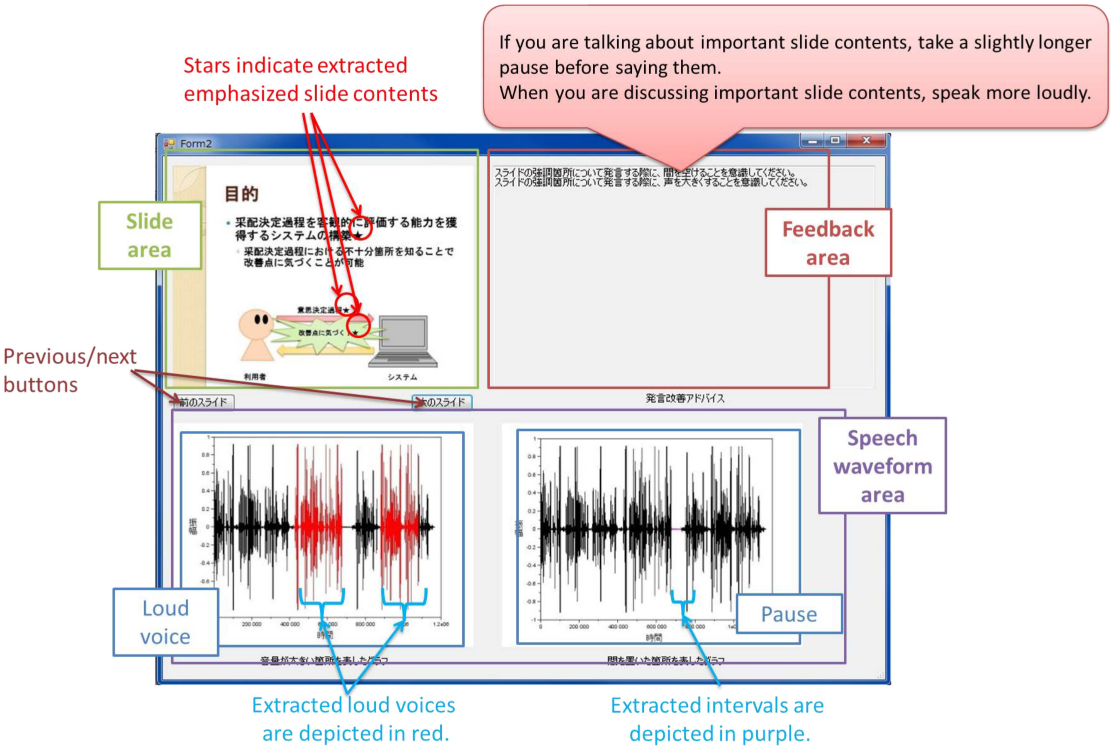
Since audiences can probably differentiate between presenters who are breathing or pausing based on the context, our system does not give feedback to the [slide] < [pause] situation.

4. Prototype System

Our prototype system gives feedback to improve the speech techniques based on the presentation slides and the sound of the presentation's speech. Scilab [14] extracts the emphasized speech parts from the sound waveform. C# is used to implement the other parts, including the user interface, the extraction mechanism of the emphasized slide intention, and the extraction mechanism of the intention differences.

The presenter selects a Microsoft PowerPoint presentation file and a speech sound file in a wav format [15]. To divide the speech sound for each slide, the speech sound is recorded for each slide and stored as different files. The file names need to be the number of slides, such as "1.wav", to correspond to the slide's sound speech file. In addition, because of Scilab's limitations, each speech sound must be shorter than one minute. This limitation might be solved when Scilab's ability to process data increases.

When the start button is pushed, the system starts analyzing the files. It creates an image of each presentation slide by attaching a star to the extracted emphasized slide contents. It also generates two images of the speech waveform; one highlights the loud parts in red and the other shows the pause parts in purple. The screenshot in Figure 4 shows the result. The Slide area shows one slide. In this area, the extracted emphasized slide contents are marked by stars. The Speech waveform area contains two images that show the same speech waveform of the slide; the left image highlights the loudness part and the right image shows the pause part. The feedback generated for this slide is shown in the Feedback area. We prepared the three types of feedback described in Section 3.2. Figure 4 gives *"If you are talking about important slide contents, make a slightly longer pause before saying them,"* and *"When you are discussing important slide contents, speak more loudly."* By pushing the Previous and Next buttons, the analysis results can be checked.



If you are talking about important slide contents, take a slightly longer pause before saying them.
When you are discussing important slide contents, speak more loudly.

スライドの強調箇所について発言する際に、間を空けることを意識してください。
スライドの強調箇所について発言する際に、声を大きくすることを意識してください。

目的

- 採配決定過程を客観的に評価する能力を獲得するシステムの構築★
  - 採配決定過程における不十分箇所を知ることで改善点に気づくことが可能

Stars indicate extracted emphasized slide contents
Form2
Slide area
Feedback area
意思決定過程★
改善点に気づく！★
利用者
システム
Previous/next buttons
前のスライド
次のスライド
発言改善アドバイス
Speech waveform area
振幅
時間
Loud voice
Pause
音量が大きい箇所を表したグラフ
間を置いた箇所を表したグラフ
Extracted loud voices are depicted in red.
Extracted intervals are depicted in purple.

Figure 4. System interface.

5. Evaluation

5.1. Experiment Overview

We experimentally evaluated the effectiveness of our system for improving presentation speeches. Our subjects were six computer science students (*A* to *F*) from our university with previous (but unsatisfactory results) presentation experience. None of the subjects belong to our research group, and they were unfamiliar with our system before the experiment. In the experiment, they presented three sets of slides about their research. We recorded their speech voices by headset and their presentation scenes by video camera. After the first presentation, they used the system and gave their presentations again. The speech and movie were also recorded again.

The types of feedback generated by the system for each subject are shown in Table 1. *Y* indicates that feedback was generated and *N* means that no feedback was given.

Table 1. Types of generated feedback for all subjects.

Subject	[Slide] > [Loudness]	[Slide] < [Loudness]	[Slide] > [Pause]
A	Y	N	N
B	Y	N	Y
C	N	Y	Y
D	Y	N	Y
E	N	Y	Y
F	Y	N	Y

5.2. Results and Discussion

Table 2 shows the opinions of the subjects from the questionnaires about the extracted emphasis-intentions and feedback from the system. For the speech waveform, showing “pause” was helpful for some subjects. However, the emphasis-intentions extracted from the slide area were inappropriate for other subjects who wanted to emphasize the texts written in the figures. Since our current system does not analyze the text in the figures, we need to devise a method that judges the importance of the text in figures. As for feedback, because our current system can generate only three types of advice, subjects tend to repeatedly get the same feedback for different slides. Since we believe that receiving identical feedback many times strongly impacts the subjects, the amount of identical feedback is not a big problem.

Table 2. Questionnaire results: extracted emphasis-intentions and feedback.

Target	**Please Give Your Opinions about the Extracted Emphasis-Intentions and Feedback of the System**
Speech waveform area	"Showing the pause parts was effective. I realized that I need to make more pauses in my speech."
Slide area	"I wanted to hear my voice while watching the speech waveform." "The extracted emphasis-intentions were slightly different from my intention."
Feedback area	"I didn't understand the criteria for extracting the emphasis-intentions." "I got the same feedback too many times."

Next we evaluated the effects of our system on presenter intentions (Table 3). Subjects answered questionnaires both before and after using it. The underlined words represent either intonations or pauses. According to the comments, only subject *A* was conscious of pausing, and subject *F* tried to add intonation during her first presentation. After using the system, subjects *A*, *B*, *C*, and *E* tried to speak with more intonation, and subjects *C* and *E* added pauses before important contents. Subjects *D* and *F* did not describe how they presented the important contents, but they did pay attention to them. Therefore, our system effectively increased awareness in the presenters of important contents. It might also change speech techniques for some presenters who might begin to emphasize critical contents.

Table 4 addresses the information given by the system that affected speech intentions. Subjects were able to select more than one answer. According to the result, the system's feedback was very effective. Although the number of answers was small, some subjects deemed the information from the slide and speech waveform areas as useful.

Table 3. Questionnaire result: intention in speech.

To What Did You Pay Attention during Your Speech?		
Subject	**After First Presentation (Before Using System)**	**After Second Presentation (After Using System)**
A	I added more pauses before important points.	I added more intonation to my voice.
B	I tried to precisely explain the words and the ideas on my slides.	As the system pointed out, I tried to use intonation to help my audience understand my intention.
C	Nothing particular	I spoke with more intonation and added pauses before the important contents.
D	Nothing particular	I concentrated on the important contents.
E	I tried to explain the important contents and the relations among them.	To emphasize the important contents, I tried to speak in a louder voice and to pause before them.
F	I tried to speak in a louder and clearer voice.	I concentrated on the importance of the contents on each slide.

Table 4. Questionnaire result: effective information given by system.

Subject	**What Information Changed Your Speech Intention?**
A	Feedback area and Speech waveform area
B	Feedback area
C	Feedback area
D	Slide area
E	Feedback area and Speech waveform area
F	Feedback area & Slide area

The subjects watched two videos of other subjects taken before and after using the system and answered questionnaires. That is, subject *A* watched subject *B*'s presentation, subject *B* watched subject *C*'s presentation, and so on. Subjects could watch the videos as many times as they wanted. In their questionnaires, subjects evaluated the presentation quality by selecting one from the following choices: "improving," "worse," and "no change." No one answered "worse" (Table 5). For the improved aspects, the subjects gave the following answers: "The presenter added pause to his presentation, which gave it more power and clarity," and "The presenter′s tone changed during the presentation, which made it more effective." These results suggest that our system can enhance the speech skills of presenters. Note that since our subjects used the system before evaluating the presentations of other subjects, they were undoubtedly influenced simply by it. To check the validity of the results shown in Table 5, we showed all of the videos to a member of our department, who found slight improvement in three of the four videos that were selected as "Improved" in Table 5. This subjective opinion suggests that Table 5′s result has some reliability. Future work will investigate with a larger audience (who did not use our system) to evaluate whether the presentations improved.

Table 5. Questionnaire result: quality change of second presentation.

How did the Second Presentation Compare with the First Presentation	The Number of Subjects
Improved	4
No change	2
Worse	0

6. Conclusions

This paper proposed a novel method for improving presentation speech techniques by representing emphasis-intentions for important slide contents. Our system extracted the amount of emphasis-intentions from PowerPoint slides and the speech and provided feedback if the amounts were different for each slide. With our system, presenters can improve their speech techniques without relying on actual audiences. Based on our experimental result, even though the number of subjects was insufficient, their presentations improved with the system. In addition, presenters became aware of adding intonation to their speeches to make them more powerful.

Currently, we define two factors, "loudness" and "pause," by which audiences can recognize the emphasis-intentions of presenters. However, since we failed to evaluate whether audiences experienced the emphasis-intentions based on these two factors, we must evaluate whether these two factors effectively grasp emphasis-intentions by investigating the audience responses.

Our current system extracts the loudness part of voices using amplitudes. Since their thresholds for detecting loudness parts such as a, b, and r may be different based on the situations, we need to heuristically set them for each presenter. To avoid such a problem, short-time energy is an alternative way to detect the loudness part of the voice. For our next step, we will introduce short-time energy as a substitute for amplitude and evaluate whether it can correctly detect the loudness part.

The final goal of our research is to determine the exact objects on which presenters are focusing during their speeches and support them to speak so that audiences can correctly identify the emphasized objects. However, in its current stage, our system only handles the number of emphasis-intentions from slides and speech and cannot determine the slide objects targeted by the emphasized speech parts. If keywords could be extracted from speech, we might determine the slide objects on which the speech is focusing. For our next challenge, we will develop a mechanism to determine the slide objects that the presenter is explaining by applying a speech recognition tool to improve the system's feedback. We will devise a method for evaluating the presentation speech techniques more precisely based on the emphasized contents: that is, detecting speech parts where the presenters emphasize unimportant contents and where presenters do not emphasize important contents.

Acknowledgments: This work was partially supported by JSPS KAKENHI Grant Numbers 25330425.

Author Contributions: Tomoko Kojiri contributed to the overall framework, system mechanisms, experiments, and the reporting. Takaya Kaji contributed to the system mechanisms, system implementation, and experiments.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Hayama, T.; Nanba, H.; Kunifuji, S. Alignment between a technical paper and presentation sheets using a hidden Markov model. In Proceedings of the International Conference on Active Media Technology, Kagawa, Japan, 19–21 May 2005; pp. 102–106.
2. Hasegawa, S.; Tanida, A.; Kashihara, A. Recommendation and Diagnosis Services with Structure Analysis of Presentation Documents. In Proceedings of the Knowledge-Based and Intelligent Information and Engineering Systems, LNCS, Kaiserslautern, Germany, 12–14 September 2011; Volume 6881, pp. 484–494.
3. Hanaue, K.; Watanabe, T. Externalization support of key phrase channel in presentation preparation. *J. Intell. Decis. Technol.* **2009**, *3*, 85–92.
4. Kojiri, T.; Iwashita, N. Presentation Speech Creation Support Based on Visualization of Slide Relations. *IEICE Trans. Inf. Syst.* **2014**, *97*, 893–900. [CrossRef]
5. Okamoto, R.; Kashihara, A. Back-Review Support Method for Presentation Rehearsal Support System. In Proceedings of the Knowledge-Based and Intelligent Information and Engineering Systems, LNCS, Kaiserslautern, Germany, 12–14 September 2011; Volume 6882, pp. 165–175.
6. TEAD: Ideas Worth Spreading. Available online: http://www.ted.com/ (accessed on 20 November 2015).
7. How to Make Presentations. Available online: http://www.kent.ac.uk/careers/presentationskills.htm (accessed on 20 November 2015).
8. Carnegie, D.; Carnegie, D. *The Quick and Easy Way to Effective Speaking*; Pocket Books: New York, NY, USA, 1990.
9. The Importance of Intonation and Rhythm. Available online: http://presentinenglish.com/the-importance-of-intonation-and-rhythm (accessed on 20 November 2015).
10. Kurihara, K.; Goto, M.; Ogata, J.; Matsuzaka, Y.; Igarashi, T. A Presentation Training System using Speech and Image Processing. In Proceedings of 14th Workshop on Interactive Systems and Software, Iwate, Japan, 6–8 December 2006; pp. 59–64. (In Japanese)
11. Yokota, H.; Kobayashi, T.; Muraki, T.; Naoi, S. UPRISE: Unified Presentation Slide Retrieval by Impression Search Engine. *IEICE Trans. Inf. Syst.* **2004**, *87*, 397–406.
12. Wang, Y.; Sumiya, K.A. Browsing Method for Presentation Slides Based on Semantic Relations and Document Structure for E-Learning. *Inf. Media Technol.* **2012**, *7*, 328–342. [CrossRef]
13. Koninga, B.B.; Tabbersa, H.K.; Rikersa, R.M.J.P.; Paas, F. Attention Cueing in an Instructional Animation: The Role of Presentation Speed. *Comput. Hum. Behav.* **2011**, *27*, 41–45. [CrossRef]
14. Scilab Enterprises. Available online: http://www.scilab.org/ (accessed on 19 November 2015).
15. WAVE Audio File Format. Available online: http://www.digitalpreservation.gov/formats/fdd/fdd000001.shtml (accessed on 19 November 2015).