Evaluation and Advice System for Improving the Manner of Speaking in Lectures Using Features of Filled Pauses

Wataru Naito* and Hiromitsu Nishizaki† and Yoshihiro Sekiguchi†
* Department of Education Interdisciplinary Graduate School of Medicine and Engineering, University of Yamanashi, Japan
E-mail: wata@alps-lab.org
† Department of Research Interdisciplinary Graduate School of Medicine and Engineering, University of Yamanashi, Japan
E-mail: hnishi@yamanashi.ac.jp, sekiguti@yamanashi.ac.jp

***Abstract*—This paper describes an evaluation and advice system for improving the manner of speaking in lectures. Our developed system evaluates the manner of speaking using decision trees. These decision trees are trained using features of filled pauses that are included in lecture speeches. Recently, it has been learned that filled pauses in speech influence the understanding of listeners and change the overall impression of the speech. Our decision-tree-based evaluation and advice system detects filled pauses using automatic speech recognition. The system then uses the features of the filled pauses in decision trees. Finally, the system evaluates the manner of speaking at three evaluation levels and provides advice to the speaker. In the evaluation experiment, our system worked well and advice given improved the manner of speaking in four out of five instances.**

I. Introduction

Improving presentation skills is important for achieving success in business. Most universities in Japan provide support for faculty development in terms of presentation in lectures so that teachers can improve the quality of academic lectures. Presentations in academic meetings are also important.

There are various factors affecting the quality of lectures, such as content, the manner of speaking, speaking skills, and the form of lectures. It is very important for us to pay attention to how we speak in order to effectively convey the correct meaning of what we speak. The manner of speaking affects the extent to which people understand our speech and listeners' influencing our listeners overall impression of our speech. In lectures, both the techniques of presentation and the manner of speaking are very important.

Our ultimate goal is to develop a speech training system that can provide appropriate advice to a speaker about his/her weak points in speech by automatically evaluating his or her manner of speaking. For that, it is necessary to capture the characteristics of speech. We have already analyzed lecture-style speech and reported the investigation results of various features of lecture speech#gn[1]. This paper, we particularly focus on filled pauses in spontaneous speech, because using filled pauses well in speech may improve speaking skills in general.

Filled pauses are a type of interjections that speakers do not intend to utter, and are sometimes used to maintain control of a speech while thinking of what to say next. Furthermore, filled pauses do not add any new information, and do not have any meaning themselves#[2]. For example: "**Um** *I'd like to talk about* **uh** *binary tree.*" In this sentence, "**Um**" and "**uh**" are filled pauses. In Japanese, "eqto," "ah," "mah," and so on, are usually used as a filled pause.

There are previous studies related to the analysis of filled pauses [3][4][5][6]; this paper aims to develop an automatic evaluation and advice system on the manner of speaking, using the characteristics of filled pauses included in lecture speech, and based on a decision tree framework. We have also reported on how filled pauses used in spontaneous speech, such as lectures, influence the understanding and listening ability of audiences#[1]. In the research, we revealed that suitable conditions of filled pauses included in lecture speech improved understanding of lecturers' talk and listening ability of audiences. This was demonstrated by performing a listening experiment in which many subjects compared an original speech with the speech that had suitable conditions of filled pauses manually that were inserted from the original speech. This result was very interesting. It suggests that practicing how to use filled pauses in speech may be one way of to becoming a good speaker.

Therefore, we attempt to develop an evaluation system that can evaluate the manner of speaking in a lecture speech by using a large vocabulary continuous speech recognition (LVCSR) system and a decision tree framework. The system produces the evaluation result in three levels. The system can also provide advice to the speaker on the basis of the clustering result form decision trees.

The system was evaluated through two experiments: one rating the manner of speaking and the other monitoring the effectiveness of advice. The results from the ratings showed that the system obtained a 40% approval rating from human test subjects. In the other experiment, advice from the system improved the manner of speaking in most speeches.

II. Effect of Filled Pauses on Listeners

As described in Section 1, we investigated how filled pauses that are included when delivering speech influence understanding and change the impression of a speech, as shown through research and experiments conducted on trial subjects#[1].

In the research, we especially focused on three characteristics: "rate," "position," and "duration" of filled pauses. Speeches, in which the characteristics of filled pauses were manually changed, were prepared for a listening experiment. Subjects listened to the speech and responded to a questionnaire related to the understanding of contents of speech and listening ability. The results of the listening experiment helped us to deduce information about each of the these characteristics of filled pauses.

Lecture speeches (original speeches) were manually converted into speeches that contained suitable conditions of filled pauses, and subjects compared the original speeches with the converted speeches on the basis of understanding and listening ability. Most of the subjects rated the converted speeches as

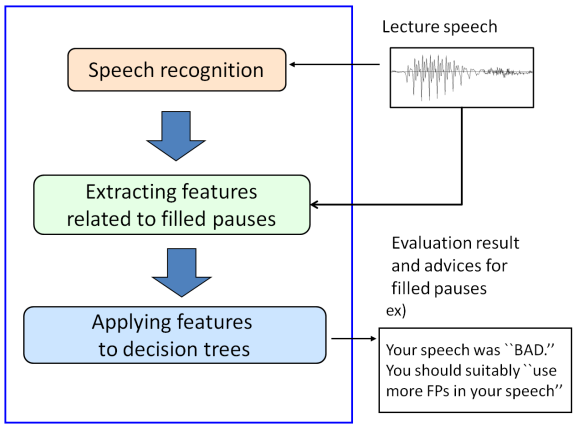

Fig. 1. Block diagram of evaluation and advice system.

being better speeches. Modifying the characteristics of filled pauses was the only change could change the impression of the listening subject. This suggests that the use of filled pauses in speech is one of the important factors in improving the manner of speaking. Therefore, we have been developing a system that can automatically evaluate how to use filled pauses in speech.

III. Evaluation and Advice System

A. System overview

Fig. 1 shows a block diagram of our evaluation and advice system. First, the system extracts features of filled pauses by performing automatic speech recognition. By applying these features to decision trees, the system can evaluate the manner of speaking of the input speech in three levels: "BAD," "NORMAL," and "GOOD,". In addition to this, the system gives some advice to the speaker.

B. Lecture speeches for training and evaluation

We used simulated lecture speeches from the Corpus of Spontaneous Japanese (CSJ)#[7] for training and testing data.

Samples of 72 speeches from CSJ were used for training decision trees, and 20 speeches were used for evaluation. Eight subjects listened to the 72 speeches that comprised the middle part of a lecture speech with a duration of 60, and answered the questionnaire[1] on the manner of speaking.

C. Speech recognition for extracting features

We used an LVCSR system called Julius[8] to detect filled pauses in speeches. First, Julius detects filled pauses, then the features of the pauses are extracted. Each speech is segmented into utterances based on a 200#ms of short pause duration.

Julius uses a language model (word trigram model) developed from CSJ, which includes a number of spontaneous lecture speeches with transcriptions. Acoustic triphone-based models (HMMs) were developed from approximately 600 h of lecture speeches in CSJ. The detection accuracy of filled pauses by Julius was more than 80%.

D. Features of filled pauses

We used 13 types of features of filled pauses (FPs) as follows:

FP ratio (number of words):
: ratio of number of FPs to all words included in a speech.

FP ratio (duration):
: ratio of total duration time of FPs to a whole speech.

Position ratio (head):
: ratio of number of HEAD FPs to all FPs in an utterance.

Position ratio (isolated):
: ratio of number of ISOLATED FPs to the total number of FPs in a whole speech.

Duration (head):
: average duration of all HEAD FPs in a speech.

Duration (middle):
: average duration of all MIDDLE FPs in a speech.

Duration rate (head):
: ratio of total duration time of all HEAD FPs to that of all FPs.

Difference of pitch (front):
: difference between the pitch of FP and that of the word in front of the FP.

Difference of pitch (back):
: difference between the pitch of FP and that of the word in back of the FP.

Difference of power (front):
: difference between the power value of FP and that of the word in front of the FP.

Difference of power (back):
: difference between the power value of FP and that of the word after the FP.

SP ratio (number of words):
: ratio of number of short pauses (SPs) to all words included in a speech.

SP ratio (duration):
: ratio of total duration time of SPs to a whole speech.

Difference in power and of pitch are averaged out for HEAD and MIDDLE FPs. These feature parameters depend on "rate," "position," and "duration" (described in Section II). In addition, we used features related to "pitch" (frequency) and "power" these are acoustic parameters.

Filled pauses are classified as follows:

HEAD FP:
: an FP located at the beginning of an utterance:
`<s>|<sp>␣FP␣word`

MIDDLE FP:
: an FP located anywhere except at the beginning of an utterance:
`word␣FP␣word|</s>|<sp>`

ISOLATED FP:
: an FP uttered in isolation:
`<s>|<sp>␣FP␣</s>|<sp>`

where "<s>" denotes the start of an utterance, "</s>" means the end of an utterance, and "<sp>" represents a short pause whose duration time is more than 200#ms. These symbols are output from Julius. The position of FPs is determined by whether these symbols are located before or after the FPs.

E. Decision tree training

We used the ADTree algorithm#[9] to evaluate ratings and give advice on the manner of speaking. Decision trees were trained by Weka#[10], an open source machine learning software.

As described in Section III-B, we used 72 speeches in CSJ to train decision trees. Each speech had been rated as being at one of the four levels by human subjects.

[1]Rating score is from 1 to 4.

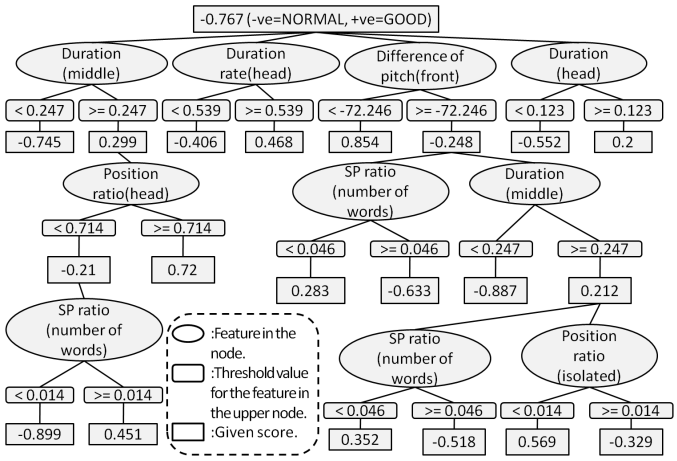

Fig. 2. The decision tree that classifies an input speech as “GOOD” or “ NORMAL.”

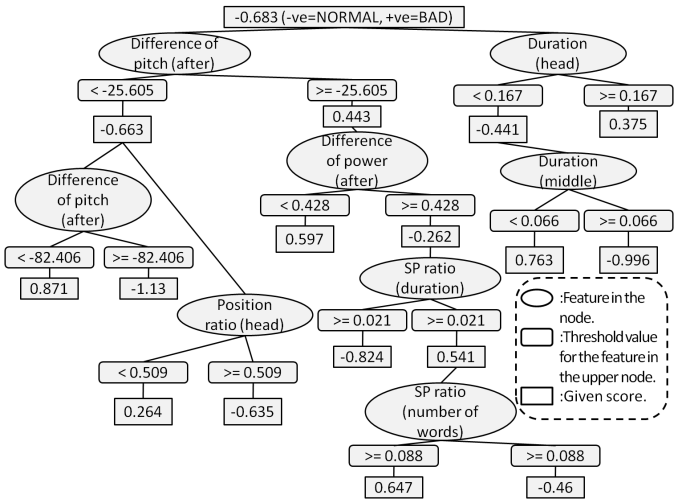

Fig. 3. The decision tree that classifies an input speech as “NORMAL” or “BAD”.

	your value of feature	threshold for getting better score	advice
Total evaluation	bad		
FP ratio (numver of words)	5.67%		
position ratio (head)	58.34%	>=71.4%	increase HEAD FPs
position ratio (isolated)	14.29%	<1.4%	decrease ISOLATED FPs
duration (head)	198.6ms	>=123ms, <167ms	slightly shorten each HEAD FP
duration (middle)	344ms	>=247ms	good
duration rate (head)	44.69%	>=53.9%	slightly increase HEAD FPs
difference of pitch (front)	-33.60mel	<-72.25mel	more intonation
difference of pitch (back)	-22.86mel	<-25.65mel	slightly less intonation
difference of power (back)	1.16	>0.428	good
SP ratio (number of words)	7.14%	<=4.6%	decrease SPs
SP ratio (duration)	3.75%	<=2.1%	slightly shorten each SPs

Fig. 4. Example of advice.

The system classifies an input speech at one of the three levels: “BAD,” “NORMAL,” and “GOOD.” For this, the system standardizes evaluation values to 1.0 of mean value and variance value. We defined speeches with standardized evaluation values of 1.0 or more as “GOOD,” and speeches with values of -1.0 and less as “BAD.” The other speeches were defined as “NORMAL.” Finally, 10 “GOOD” speeches, 50 “NORMAL” speeches and 12 “BAD” speeches were prepared.

We trained the two types of decision trees by Weka. Figs. 2 and 3 show the trained trees. One of them might classify a speech as “GOOD” or “NORMAL.” The other might classify a speech as “NORMAL” or “BAD.” The final evaluation is determined on the basis of the voting of the evaluation of the two trees. If one of the trees classifies a speech as “GOOD” and the other tree classifies it as “NORMAL,” then the final evaluation is “GOOD.” If one tree classifies the speech as “NORMAL” and the other as “BAD,” then the final evaluation is “BAD.” For all other combinations, the final evaluation is “NORMAL.”

A score can be obtained for every node of a tree. The system provides advice using these scores. If a speech obtains a “BAD” evaluation, the system highlights bad points of the speech, and suggests how to improve the speech by tracking back to the upper node, which is the critical branch point at which the evaluation can be improved. Such advice can improve the speaker’s manner of speaking. Fig. 4 shows an example of advice. If a speech is evaluated as being “NORMAL” with a score of -0.329 (the bottom rightmost terminal node in Fig. 2), taking notice of “position ratio (isolated)” may improve the speech. In this case, the system may give the advice: “You should be aware of isolated filled pauses.” If the speaker follows the advice, the evaluation score will be improved to 0.569.

IV. System evaluation

The system was evaluated using two experiments: classification of the manner of speaking and effectiveness of the advice.

A. Rating experiment

In this experiment, the manner of speaking was evaluated using decision trees. Decision trees were trained by 72 speeches and evaluated by 20 evaluation speeches that were not used for training decision trees.

Tables I and II show the rating results represented as the classification matrices for the training (closed) and evaluation (open) speeches, respectively.

In Table I (the evaluation of closed speeches), the evaluation ratings of 71 speeches are consistent with the human subjects’ evaluations; (the agreement rate is 98.6%). On the other hand, in the open test, the system achieved on 40% agreement rate with the subjects evaluations, as shown in Table II.

Although the number of critical errors is small, it is not good result. It is a critical error if a speech evaluated as being “BAD” or “NORMAL” by the subject is classified as “GOOD” by the system. A speech receiving the “GOOD” evaluation cannot bi provided with any advice, even if the speech is really, a “BAD” speech. This is undesirable for a speaker. In the evaluation speeches, there were only just three critical errors. If these critical errors were eliminated then the system would rate 85% of the evaluation speeches as being ”GOOD.”

B. Effectiveness of advice

We also examined the effectiveness of advice provided by the system.

First, we prepared seven speeches spoken by one of authors. Next, the seven speeches were automatically evaluated and rated by the system. Then, the system gave advice to the speaker. Finally, the speaker repeated the same speeches,

TABLE I
EVALUATION MATRIX FOR CLOSED SPEECHES.

		Evaluation by the system		
		GOOD	NORMAL	BAD
Evaluation by humans	GOOD	10	0	0
	NORMAL	0	50	0
	BAD	0	1	11

TABLE II
EVALUATION MATRIX FOR OPEN SPEECHES.

		Evaluation by the system		
		GOOD	NORMAL	BAD
Evaluation by humans	GOOD	2	2	0
	NORMAL	**2**	6	5
	BAD	**1**	2	0

following the advice provided. Nine subjects listened to the original speeches, and then to the speeches following the adoption of the system as advices. The subjects were asked a question: "Which speech is better in terms of the manner of speaking?"

Five of the seven speeches improved the system's rating, as shown in Table III. In Table III, the column labeled "speech:#X" refers to the evaluation result of the speeches that were improved following the system's advice. The column labeled "speech:#Y" shows the evaluation result of the speeches that were not altered following advice. #X>#Y refers to the number of subjects who chose "speech:#X" as a speech better than "speech:Y."

Moreover, the Table III shows that four of the five speeches were improved as a result of the advice received, according to the evaluations of the manner of speaking provided by human subjects. In the case of the two speeches where the rating of the system did not change, it seems that there was no significant difference between both the speeches, with and without the advice.

These results suggest that our evaluation and advice system worked well on the test set, and seems to be effective for speakers in terms of advice given for improving the manner of speaking in lectures.

TABLE III
EFFECTIVENESS OF ADVICES PROVIDED BY THE SYSTEM.

	System evaluation		Human evaluation	
No.	speech: X	speech: Y	X>Y	X≤Y
1	GOOD	NORMAL	**8**	1
2	GOOD	NORMAL	**7**	2
3	GOOD	NORMAL	2	7
4	GOOD	BAD	**8**	1
5	GOOD	BAD	**7**	2
6	NORMAL	NORMAL	3	6
7	NORMAL	NORMAL	4	5

X>Y: The number of subjects who chose X.

V. CONCLUSIONS

In this paper, we have described the development of a system for evaluating the manner of speaking in a speech. Our system not only evaluates the speech but also gives teh speakers some advice for improving the manner of speaking. Our system evaluated speech and output advice using decision trees. Decision trees were trained using features of filled pauses manually inserted in speeches.

Our system has been able completely rate 40% of the test speeches and provide an outline evaluation for 17 of the 20 speeches; the remaining three speeches were irrelevantly classified. Most of the advice provided by the system was shown to be pertinent. By following the advice, the manner of speaking was improved.

We used speeches that were in #60s duration. In future work, we intend to refine the system by using not only the characteristics of filled pauses but also their acoustic properties. In particular, we will consider the properties that are using derived speech recognizers#[11].

ACKNOWLEDGEMENTS

This research was partially supported by the Ministry of Education, Science, Sports and Culture of Japan, Grant-in-Aid for Young Scientists, 21700807, 2009-2010.

REFERENCES

[1] M.Somiya, K.Kobayashi, H.Nishizaki, and Y.Sekiguchi, "The Effect of Filled Pauses in a Lecture Speech on Impressive Evaluation of Listeners", INTERSPEECH 2007, pp.2673–2676, 2007.

[2] A. Batliner and A. Kießling and S. Burger and E. Nöth, "Filled Pauses in Spontaneous Speech", Proc. of the 13th International Congress of Phonetic Sciences(ICPhS95), Vol.3, pp.472–475, 1995

[3] Monique E. van Donzel and Florien J. Koopmans-van Beinum, "Pausing Strategies in Discourse in Dutch", Proc. of ICSLP'96, Vol.2, pp.1029–1032, 1996

[4] Marc Swerts, Anne Wichmann and Robbrt-Jan Beun, "Filled Pauses as Markers of Discourse Structure", Proc. of ICSLP'96, Vol.2, pp.1033–1036, 1996

[5] Michiko Watanabe, Keiichi Hirose, Yasuharu Den and Nobuaki Minematsu, "Filled Pauses as Cues to the Complexity of Following Phrases", Proc. of INTERSPEECH 2005, pp. 37–40, 2005

[6] Helena Moniz, Ana Isabel Mata and M.Ceu Viana, "On Filled-Pauses and Prolongations in Eurpean Portuguese", Proc. of INTERSPEECH 2007, 2007

[7] Kikuo Maekawa, "Corpus of Spontaneous Japanese: Its Design and Evaluation", Proc. of the ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition(SSPR2003), 2003

[8] A. Lee, T. Kawahara and K. Shikano, "Julius — an open source real-time large vocabulary recognition engine", Proc. of EUROSPEECH, pp. 1691–1694, 2001

[9] Y. Freund, and L. Mason, "The Alternating Decision Tree Algorithm", ICML, pp.124–133, 1999.

[10] G. Holmes, A. donkin, and I.H. Witten, "WEKA: A Machine Learning Workbench", 2nd. AUS&NZ Conf. on Int. Info. Sys., 1994.

[11] K.Kobayashi, H.Nishizaki, and Y.Sekiguchi, "Is a speech recognizer useful for characteristic analysis of classroom lecture speech?", INTERSPEECH 2008, pp.1341–1344, 2008.