A METHOD TO ENHANCE INFORMATIZED CAPTION FROM IBM WATSON API USING SPEAKER PRONUNCIATION TIME-DB

Yong-Sik Choi, YunSik Son and Jin-Woo Jung

Department of Computer Science and Engineering, Dongguk University, Seoul, Korea

ABSTRACT

IBM Watson API is a kind of speech recognition API system which can automatically generate not only recognized words from voice signal but also generate speaker ID and timing information of each words including the starting time and the ending time. The performance of IBM Watson API is very good at the well-recorded voice signal by the clearly speaking trained speakers but the performance is not enough good when there are some noises in the recorded voice signal. This situation is easily found with movie sounds that include not only speaking voice signal but also background music or special sound effects. This paper deals with a novel method to enhance this informatized caption from IBM Watson API to resolve this noisy signal problem based on speaker pronunciation time-DB. To do this, the proposed method uses the original caption information as an additional input. By comparing the original caption with the output of IBM Watson API, the error words could be automatically detected and correctly modified. And using the speaker pronunciation time-DB containing the average pronunciation time of each word for each speaker, the timing information of each error word could be estimated. In this way, more precisely enhanced informatized captions could be generated based on the IBM Watson API. The usefulness of the proposed method is verified with two case studies with noisy voice signals.

KEYWORDS

Informatized caption, Speaker Pronunciation Time, IBM Watson API, Speech to Text Translation

1. INTRODUCTION

By the waves of 4th industrial revolution, artificial intelligence becomes one of the most promising technologies nowadays. There are so many research areas and research results from artificial intelligence. One of them is natural language processing by speech recognition. Typical speech recognition technologies include speech to text conversion.Among captions in which speech is converted into characters, captions including timing information and speaker ID information are referred to as informatizedcaptions [1, 2]. Such an informatized caption could be generated by using IBM Watson API [3]. However, the IBM Watson API is more susceptible to clipping errors due to poor recognition results when there aresome noises in the voice signal. And this situation is easily found with movie sounds that include not only speaking voice signal but also include background music or special sound effects.In order to solve this noisy voice problem, there has been a method of predicting the timing information of informatized caption based on a linear estimation formula proportional to the number of alphabets used in each word [2]. But, this linear estimation method based on the number of alphabets is not good enough when there are some silent syllables. Therefore, a novel method to enhance the informatized caption from IBM Watson API is proposed in this paperbased on the speaker pronunciation time-DB.

DOI: 10.5121/ijnlc.2018.7101

2. Speaker Pronunciation Time-Db (Spt-Db)

2.1. Structure

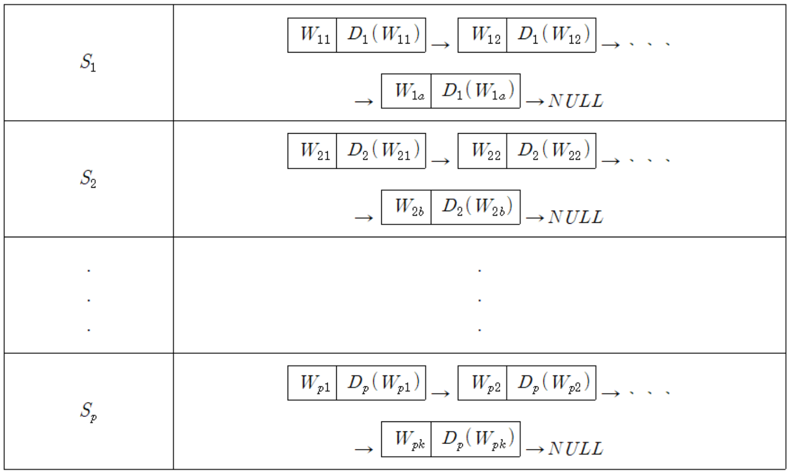

S_1
W_{11} $D_1(W_{11})$ → W_{12} $D_1(W_{12})$ → . . .
→ W_{1a} $D_1(W_{1a})$ → NULL
S_2
W_{21} $D_2(W_{21})$ → W_{22} $D_2(W_{22})$ → . . .
→ W_{2b} $D_2(W_{2b})$ → NULL
S_p
W_{p1} $D_p(W_{p1})$ → W_{p2} $D_p(W_{p2})$ → . . .
→ W_{pk} $D_p(W_{pk})$ → NULL

Figure 1. Structure of SPT-DB

SPT-DB consists of each node for each speaker(S_p) as shown in Fig. 1.The nodes consist of the average pronunciation times(D_p) of each word(W_{pk}).The nodes of the speaker are arranged in ascending order based on the average pronunciation time, and are connected to each other, and a null value is present at the end. When SPT-DB searches for a word spoken by the speaker, it searches based on the pronunciation time.

2.2. Assumption

Before proceeding with the study, the following assumptions are based on SPT-DB. [Assumption]SPT-DB is already configured for each speaker.

3. Proposed Algorithm

3.1. Algorithm Modifying Incorrectly Recognized Word Based On SPT-DB

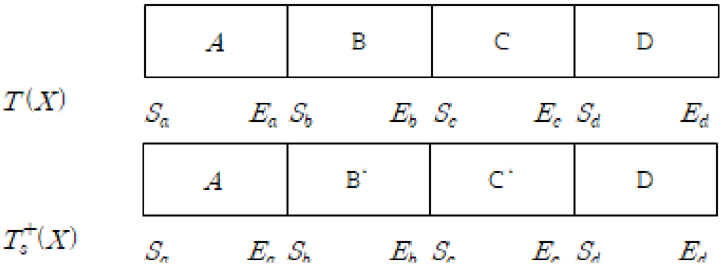

A
B
C
D
$T(X)$
S_a E_a S_b E_b S_c E_c S_d E_d
A
B'
C'
D
$T_s^+(X)$
S_a E_a S_b E_b S_c E_c S_d E_d

Figure 2. Original caption T(X) and informatized caption $T_s^+(X)$

Basically, original caption,T(X), and informatized captionfrom speech recognition result,$T_s^+(X)$, are input together.

Here, S_x and E_x mean the start time and end time of pronunciation for the word X, respectively.

[Step 1] Judge whether there is an incorrectly recognized word by comparing T (X) with $T_s^+(X)$.If there is no incorrectly recognized word, it terminates. If there is an incorrectly recognized word, go to the next step.

[Step2] Judge whether there are several consecutive words in the sequence and pass the parameter to the case.

[Step3] Modify the words in the SPT-DB based on the start and end points of the cases.

[Step4] If there is an incorrectly recognized word in the following word, repeat steps 1 to 3 and terminate if there is no incorrectly recognized word.

3.2. Case 1: There Is Only One Incorrectly Recognized Word.

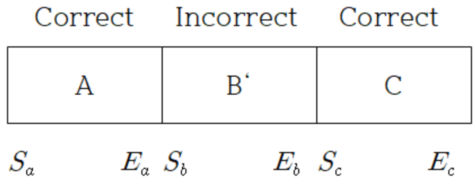


Figure3. There is one incorrectly recognized word

[Step1] Find the point at which the signal of a specific volume(dB) T or more starts for E_a to S_c and determine S_b.

[Step2] If there is a minimum time t'in $S_b$toS_c at which the signal intensity falls below a certain volume T and then remains below T until S_c, E_b = t' is determined. If there is no t'satisfying the above condition, $E_b = S_c$.

[Step3]Returns the start time and end time.

3.3. Case 2: There Are Two Incorrectly Recognized Word.

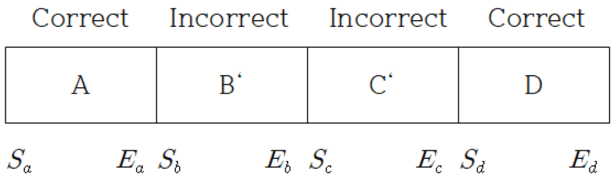


Figure 4. Two incorrectly recognized word

[Step1] Find the point at which the signal of a specific volume(dB) T or more starts for E_a to S_c and determine S_b.

[Step2] If there is a minimum time t'in $S_b$toS_c at which the signal intensity falls below a certain volume T and then remains below T until S_c, E_b = t' is determined. If there is no t'satisfying the above condition,$E_b = S_c$.

[Step3] The ending point of the current word is obtained by multiplying the start time of the current word by the ratio of the pronunciation time of the two words to the average pronunciation time of the current word.The following are summarized as follows.

$$E_b = S_b + (E_c - S_b) \times \frac{D(S,B)}{D(S,B) + D(S,C)}$$

[Step4] Returns the start time and end time.

3.4. Case 3: There Are More Than Three Incorrectly Recognized Word.

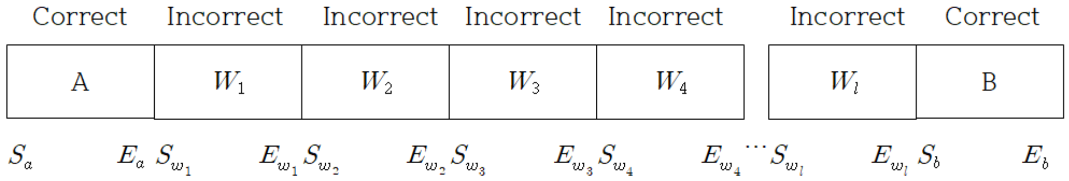


Figure 5. More than three incorrectly recognized word

[Step1] Find the point at which the signal of a specific volume(dB) T or more starts for E_a to S_{w2} and determine S_{w1}.

[Step2] If there is a minimum time t'in $S_{w1}$to S_{w2} at which the signal intensity falls below a certain volume T and then remains below T until S_{w2}, E_{w1} = t' is determined. If there is no t'satisfying the above condition,$E_{w1} = S_{w2}$.

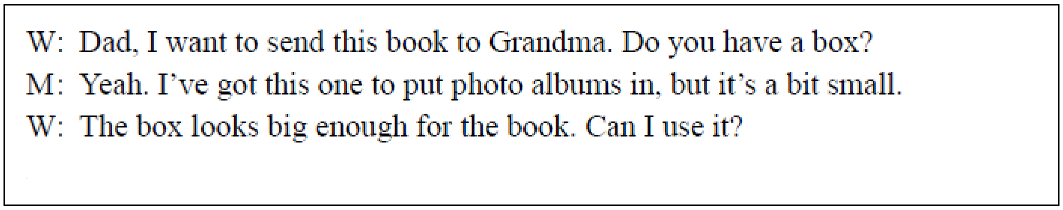
[Step3]The ending point of the current word is obtained by multiplying the start time of the current word by the ratio of the pronunciation time of the incorrectly recognized words to the average pronunciation time of the current word.The following are summarized as follows.

$$E_{w_i} = \left(E_{w_l} - S_{w_1}\right) \times \frac{D(W_i)}{\sum_{i=1}^{l} D(W_i)}$$

[Step4] Returns the start time and end time.

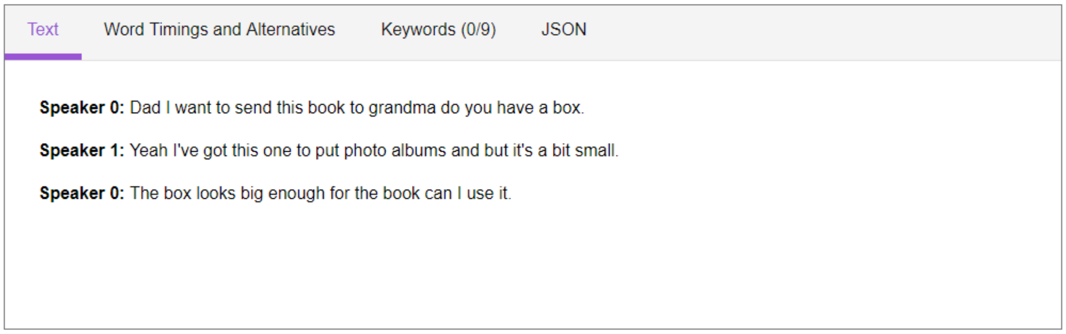
4. Case Study Ⅰ

The case was tested based on English listening assessment data. Fig. 6 shows a problem of the English listening evaluation for university entrance examination. Fig. 7 and Fig. 8 show the result of speech recognition using the IBM Watson API.Table1 and Table2 list the time information of the caption at that time, it is expressed as [start time–end time].Using the IBM Watson API, speech recognition in an environment with no constraints results in high accuracy as shown in Fig. 7. Table 1shows that incorrectly recognition word is (B, 10) and the accuracy of speech recognition is 97.56%.However, in a noisy environment like Fig.8, the accuracy dropped significantly. Table 2shows that incorrectly recognition words are(A, 1), (A, 7), (A, 7), (A, 13), (B, 2), (B, 3), (B, 4), (B, 4), (B, 9), (B, 10) and (C, 3), and the accuracy of speech recognition is 73.17%.For reference, the original voice source was synthesized with raining sound using Adobe Audition CC 2017 to create a noisy environment. If we improve the proposed algorithm with noise, we can obtain the same result as Fig.9 and Table3. The accuracy of speech recognition is 100% by the help of original caption and each word includes its own start time and end time.



W: Dad, I want to send this book to Grandma. Do you have a box?
M: Yeah. I’ve got this one to put photo albums in, but it’s a bit small.
W: The box looks big enough for the book. Can I use it?

Figure 6. Original caption of Case Study Ⅰ



Speaker 0: Dad I want to send this book to grandma do you have a box.

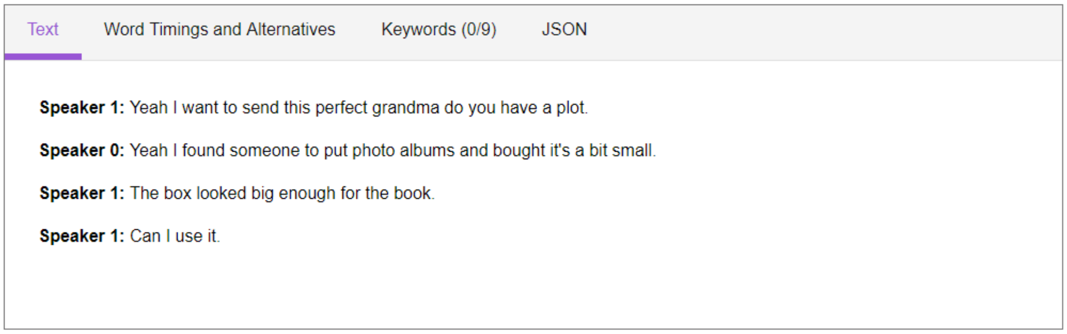
Speaker 1: Yeah I've got this one to put photo albums and but it's a bit small.

Speaker 0: The box looks big enough for the book can I use it.

Figure 7. Recognition of original voice without noise by IBM Watson system

Table 1. Informatized caption from original voice without noise by IBM Watson system

ord Sentence		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	Speaker 0	Dad	I	want	to	send	this	book	to	grandma	do	you	have	a	box	
		0.03-0.58	0.74-0.87	0.87-1.19	1.19-1.35	1.35-1.66	1.66-1.83	1.83-2.1	2.1-2.24	2.24-2.89	3.21-3.45	3.45-3.67	3.67-3.95	3.95-4.03	4.03-4.75	
B	Speaker 1	Yeah	I’ve	got	this	one	to	put	photo	albums	and	but	it’s	a	bit	small
		5.22-5.7	6.01-6.27	6.27-6.62	6.62-6.86	6.86-7.15	7.15-7.26	7.26-7.48	7.48-7.88	7.88-8.29	8.29-8.59	8.69-9.1	9.28-9.48	9.48-9.55	9.55-9.81	9.81-10.51
C	Speaker 0	The	box	looks	big	enough	for	the	book	Can	I	use	it			
		10.86-10.99	10.99-11.41	11.41-11.67	11.67-11.96	11.96-12.26	12.26-12.47	12.47-12.6	12.6-13.16	13.46-13.71	13.71-13.79	13.79-14.12	14.12-14.42			



Speaker 1: Yeah I want to send this perfect grandma do you have a plot.

Speaker 0: Yeah I found someone to put photo albums and bought it's a bit small.

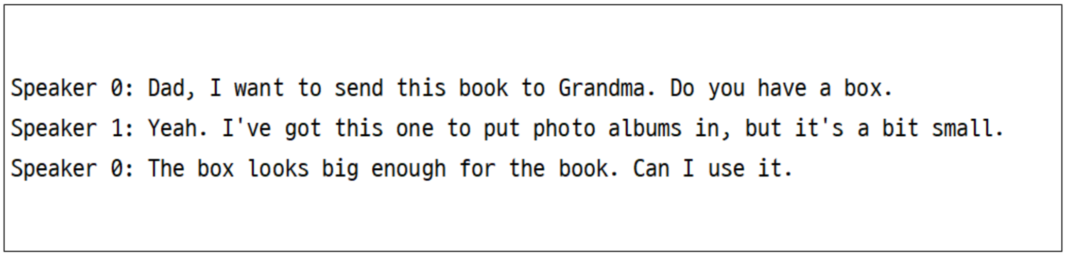
Speaker 1: The box looked big enough for the book.

Speaker 1: Can I use it.

Figure 8. Recognition of mixed voice with rain noise by IBM Watson system

Table 2. Informatized caption from mixed voice with rain noise by IBM Watson system

Word Sentence		1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	Speaker 1	Yeah	I	want	to	send	this	perfect	grandma	do	you	have	a	plot	
		0.05-0.46	0.74-0.87	0.87-1.19	1.19-1.35	1.35-1.66	1.66-1.82	1.82-2.29	2.29-2.88	3.32-3.44	3.44-3.67	3.67-3.96	3.96-4.02	4.02-4.37	
B	Speaker 0	Yeah	I	found	someone	to	put	photo	albums	and	bought	it's	a	bit	small
		5.27-5.66	6.08-6.21	6.21-6.55	6.77-7.15	7.15-7.25	7.25-7.48	7.55-7.87	7.87-8.29	8.29-8.53	8.83-9.08	9.29-9.48	9.48-9.55	9.55-9.81	9.81-10.27
C	Speaker 1	The	box	looked	big	enough	for	the	book						
		10.83-10.98	10.98-11.41	11.41-11.71	11.71-11.92	11.92-12.25	12.25-12.47	12.47-12.59	12.59-13.03						
D	Speaker 1	Can	I	use	it										
		13.52-13.7	13.7-13.79	13.79-14.12	14.12-14.29										



Speaker 0: Dad, I want to send this book to Grandma. Do you have a box.
Speaker 1: Yeah. I've got this one to put photo albums in, but it's a bit small.
Speaker 0: The box looks big enough for the book. Can I use it.

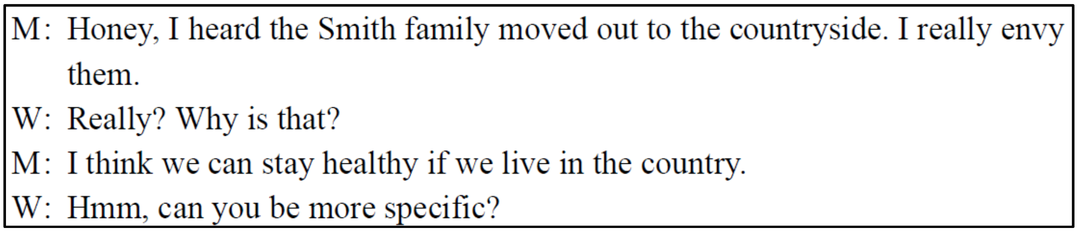
Figure 9. Speech recognition result modified by proposed algorithm

Table3.Informatized caption modified by the proposed algorithm

ord Sentence		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	Speaker 0	Dad	I	want	to	send	this	book	to	Grandma	Do	you	have	a	box	
		0.03-0.58	0.74-0.87	0.87-1.19	1.19-1.35	1.35-1.66	1.66-1.83	1.83-2.1	2.1-2.24	2.24-2.89	3.21-3.45	3.45-3.67	3.67-3.95	3.95-4.03	4.03-4.75	
B	Speaker 1	Yeah	I've	got	this	one	to	put	photo	albums	in	but	it's	a	bit	small
		5.22-5.7	6.01-6.27	6.27-6.62	6.62-6.86	6.86-7.15	7.15-7.26	7.26-7.48	7.48-7.88	7.88-8.29	8.29-8.59	8.69-9.1	9.28-9.48	9.48-9.55	9.55-9.81	9.81-10.51
C	Speaker 0	The	box	looks	big	enough	for	the	book	Can	I	use	it			
		10.86-10.99	10.99-11.41	11.41-11.67	11.67-11.96	11.96-12.26	12.26-12.47	12.47-12.6	12.6-13.16	13.46-13.71	13.71-13.79	13.79-14.12	14.12-14.42			

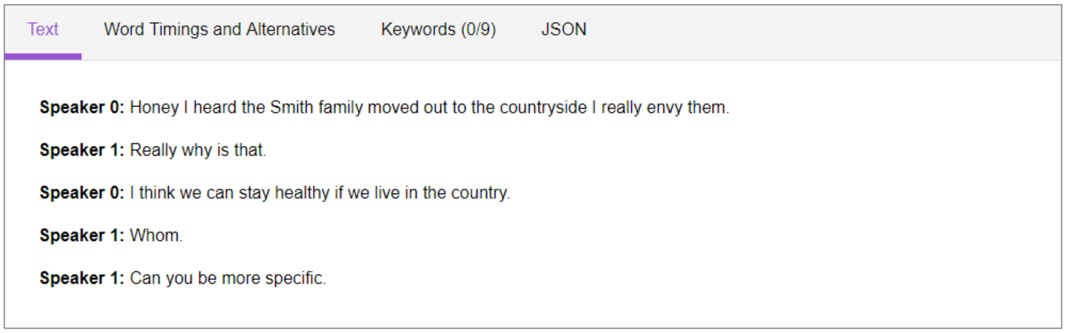
5. CASE STUDYⅡ

The case was tested based on English listening assessment data. Fig.10 shows a problem of the English listening evaluation for university entrance examination. Fig.11 and Fig. 12 show the result of speech recognition using the IBM Watson API.Table 4 and Table 5 list the time information of the caption at that time, it is expressed as [start time–end time].Using the IBM Watson API, speech recognition in an environment with no constraints results in high accuracy as shown in Fig.11. Table 4shows that incorrectly recognized word is (D, 1) and the accuracy of speech recognition is 97.29%.However, in a noisy environment like Fig.12, the accuracy dropped significantly. Table 5shows that incorrectly recognized words are(A, 1), (A, 11)and (A, 11), and the accuracy of speech recognition is91.89%.For reference, the original voice source was synthesized with raining sound using Adobe Audition CC 2017 to create a noisy environment.If we improve the proposed algorithm with noise, we can obtain the same result as Fig.13 and Table 6. The accuracy of speech recognition is 100% by the help of original caption and each word includes its own start time and end time.



M: Honey, I heard the Smith family moved out to the countryside. I really envy them.
W: Really? Why is that?
M: I think we can stay healthy if we live in the country.
W: Hmm, can you be more specific?

Figure 10. Original caption of Case StudyⅡ



Speaker 0: Honey I heard the Smith family moved out to the countryside I really envy them.

Speaker 1: Really why is that.

Speaker 0: I think we can stay healthy if we live in the country.

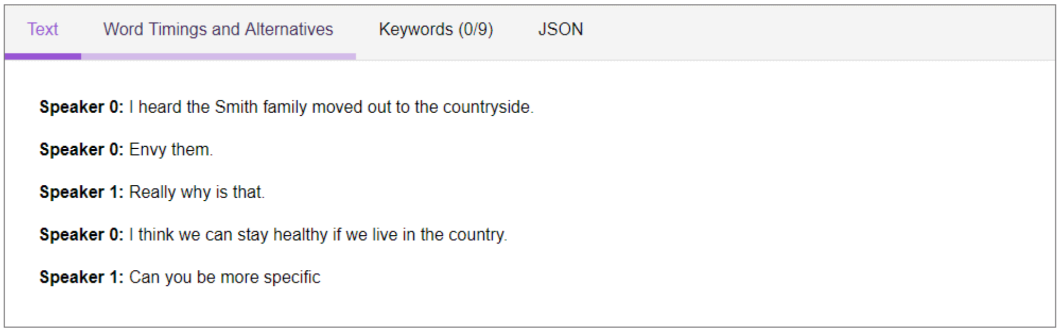
Speaker 1: Whom.

Speaker 1: Can you be more specific.

Figure 11. Recognition of original voice without noise by IBM Watson system

Table 4. Informatized caption from original voice without noise by IBM Watson system

	ord Sentence	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	Speaker 0	Honey	I	heard	the	Smith	family	moved	out	to	the	countryside	I	really	envy	them
		0.09-0.63	0.82-1.01	1.01-1.29	1.29-1.41	1.41-1.82	1.82-2.19	2.19-2.61	2.61-2.80	2.80-2.90	2.90-3.03	3.03-4.00	4.26-4.45	4.45-4.98	4.98-5.41	5.41-5.82
B	Speaker 1	Really	why	is	that											
		6.20-6.85	7.19-7.45	7.45-7.65	7.65-8.06											
C	Speaker 0	I	think	we	can	stay	healthy	if	we	live	in	the	country			
		8.52-8.73	8.73-9.06	9.06-9.22	9.22-9.41	9.41-9.66	9.66-10.17	10.17-10.33	10.33-10.53	10.53-10.75	10.75-10.86	10.86-10.97	10.97-11.62			
D	Speaker 1	Whom	Can	you	be	more	specific									
		12.01-12.57	13.10-13.33	13.33-13.44	13.44-13.65	13.65-13.90	13.90-14.71									



Speaker 0: I heard the Smith family moved out to the countryside.

Speaker 0: Envy them.

Speaker 1: Really why is that.

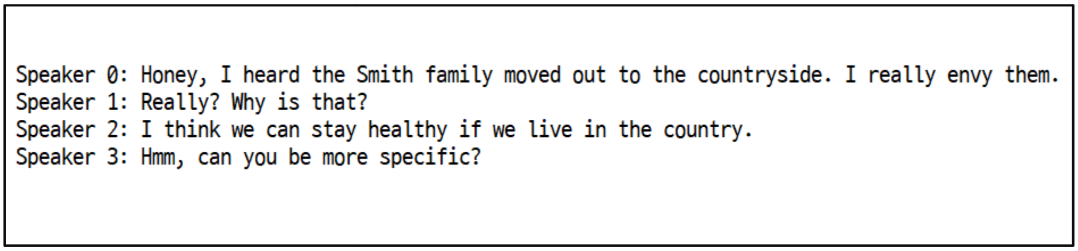
Speaker 0: I think we can stay healthy if we live in the country.

Speaker 1: Can you be more specific

Figure 12. Recognition of mixed voice with rain noise by IBM Watson system

Table 5. Informatized caption from mixed voice with rain noise by IBM Watson system

ord Sentence		1	2	3	4	5	6	7	8	9	10	11	12
A	Speaker 0	I	heard	the	Smith	family	moved	out	to	the	countryside	envy	them
		0.82-1.01	1.01-1.29	1.29-1.40	1.40-1.8	1.80-2.19	2.19-2.60	2.60-2.80	2.80-2.90	2.90-3.02	3.02-3.93	5.03-5.41	5.41-5.69
B	Speaker 1	Really	why	is	that								
		6.27-6.79	7.19-7.45	7.45-7.66	7.66-8.00								
C	Speaker 0	I	think	we	can	stay	healthy	if	we	live	in	the	country
		8.5-8.73	8.73-9.06	9.06-9.21	9.21-9.40	9.40-9.64	9.64-10.14	10.14-10.33	10.33-10.53	10.53-10.75	10.75-10.86	10.86-10.97	10.97-11.55
D	Speaker 1	Can	you	be	more	specific							
		13.12-13.33	13.33-13.44	13.44-13.65	13.65-13.90	13.90-14.66							



```
Speaker 0: Honey, I heard the Smith family moved out to the countryside. I really envy them.
Speaker 1: Really? Why is that?
Speaker 2: I think we can stay healthy if we live in the country.
Speaker 3: Hmm, can you be more specific?
```

Figure 13. Speech recognition result modified by proposed algorithm

Table 6. Informatized caption modified by the proposed algorithm

ord Sentence		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
A	Speaker 0	Honey	I	heard	the	Smith	family	moved	out	to	the	countryside	I	really	envy	them
		0.09-0.63	0.82-1.01	1.01-1.29	1.29-1.41	1.41-1.82	1.82-2.19	2.19-2.61	2.61-2.80	2.80-2.90	2.90-3.03	3.03-4.00	4.26-4.45	4.45-4.98	4.98-5.41	5.41-5.82
B	Speaker 1	Really	why	is	that											
		6.20-6.85	7.19-7.45	7.45-7.65	7.65-8.06											
C	Speaker 0	I	think	we	can	stay	healthy	if	we	live	in	the	country			
		8.52-8.73	8.73-9.06	9.06-9.22	9.22-9.41	9.41-9.66	9.66-10.17	10.17-10.33	10.33-10.53	10.53-10.75	10.75-10.86	10.86-10.97	10.97-11.62			
D	Speaker 1	Hmm	Can	you	be	more	specific									
		12.01-12.55	13.10-13.33	13.33-13.44	13.44 13.65	13.65-13.90	13.90-14.71									

6. CONCLUSIONS

In this paper, a novel method to enhance the informatized caption from IBM Watson API based on speaker pronunciation time-DB is addressed to find and modifyincorrectly recognized words from IBM Watson API. SPT-DBcontains the average pronunciation time information of each word foreach speaker and is used to correct the errors in the informatized caption obtained through the IBM Watson API. The usefulness of the proposed method is verified with two case studies with noisy voice signals.However, the proposed algorithm also has some limitations such as that SPT-DB should be created in advance because it is assumed that the information of the corresponding words already exists in SPT-DB. Furtherstudy will be conducted to modify incorrectly recognized words while performing speech recognition and simultaneously to update the SPT-DB in real time.

ACKNOWLEDGEMENTS

Following are results of a study on the "Leaders in INdustry-university Cooperation+" Project, supported by the Ministry of Education and National Research Foundation of Korea,and partially supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2015R1D1A1A09061368),and also partially supported bythis work was supported by the Korea Institute for Advencement of Technology(KIAT) grant funded by the Korean government(Motie:Ministry of Trade, Industry&Energy, HRD Program for Embedded Software R&D) (No. N001884).

REFERENCES

[1] CheonSun Kim, "Introduction to IBM Watson with case studies." Broadcasting and Media Magazine, Vol.22, No. 1,pp24-32.

[2] Yong-Sik Choi, Hyun-Min Park, Yun-Sik Son and Jin-Woo Jung, “Informatized Caption Enhancement based on IBM Watson API,” Proceedings of KIIS Autumn Conference 2017, Vol. 27, No. 2, pp105-106.

[3] IBM Watson Developer’s Page, https://www.ibm.com/watson/developer

AUTHORS

Yong-Sik Choi has been under M.S. candidate course at Dongguk university, Korea, since 2017. His current research interests include machine learning and intelligent human-robot interaction.

YunSik Son received the B.S. degree from the Dept. of Computer Science and Engineering, Dongguk University, Seoul, Korea, in 2004, and M.S. and Ph.D. degrees from the Dept. of Computer Science and Engineering, Dongguk University, Seoul, Korea in 2006 and 2009, respectively. He was a research professor of Dept. of Brain and Cognitive Engineering, Korea University, Seoul, Korea from 2015-2016. Currently, he is an assistant professor of the Dept. of Computer Science and Engineering, Dongguk University, Seoul, Korea. Also, His research areas include secure software, programming languages, compiler construction, and mobile/embedded systems.

Jin-Woo Jung received the B.S. and M.S. degrees in electrical engineering from Korea Advanced Institute of Science and Technology (KAIST), Korea, in 1997 and 1999, respectively and received the Ph.D. degree in electrical engineering and computer science from KAIST, Korea in 2004. Since 2006, he has been with the Department of Computer Science and Engineering at Dongguk University, Korea, where he is currently a Professor. During 2001~2002, he worked as visiting researcher at the Department of Mechano-Informatics, University of Tokyo, Japan. During 2004~2006, he worked as researcher in Human-friendly Welfare Robot System Research Center at KAIST, Korea. During 2014, he worked as visiting scholar at the Department of Computer and Information Technology, Purdue University, USA. His current research interests include human behaviour recognition, multiple robot cooperation and intelligent human-robot interaction.