https://doi.org/10.2352/issn.2168-3204.2018.1.0.20

Research on Applying Speech Recognition for Audio-Visual Records at the National Archives of Korea

Jae-Pyeong Kim, Yong-Min Shin, Sang-Kook Kim; National Archives of Korea; Seongnam, Gyeonggi-do, Republic of Korea

Abstract

Speech recognition technology can help searching and understanding the contents of audio-visual records that archives hold. But old video records sometimes do not guarantee good recognition results due to low signal quality or lack of vocabulary used at that time.

This paper shows actual experimental results and trials to enhance the accuracy using speech recognition toolkit based on deep learning, by training with relevant corpus data for video records in the 1950s and 1970s.

This paper also proposes a strategy for records management applications, considering of accuracies and service purposes for the future.

Introduction

Numerous audio-visual records have been transferred to the National Archives of Korea from each government agency throughout a half-century. Unlike paper or electronic documents, audio-visual records have difficulty when searching their contents with only 'title' information. Inefficient and expensive work can be performed to describe text contents or to make subtitles.

The National Archives of Korea is researching the application of speech recognition technology for audio-visual records we hold, to enhance searching efficiency and to help description work.

We used speech recognition technology based on deep neural networks with language model adaptation and experimented on old video records with several conditions.

Table 1. Collection status for audio-visual records in the National Archives of Korea, 2016

Total (items)	Movie Films	Video Tapes	Audio Tapes	Digital Files
81,318	22,464	28,534	26,170	4,150

Use of Speech to Text Technology

The objectives of our research and experiments are as follows:

a) Researching the possibility for speech to text conversion technique to apply to records and archives management

b) Applying to actual audio-visual records and analyzing the results

c) Enhancing recognition accuracy by using language model adaptation

We built a testbed system to test the effect of the new language model for old broadcasting videos.

- S/W: Korean Speech Recognizer Toolkit developed by ETRI (Electronics and Telecommunications Research Institute); a government funded research institution in Korea

- H/W: Xeon E5 CPU / 128GB MEM / GTX1070 for deep learning acceleration/ Linux CentOS

As for speech recognition toolkits, there are some commercial products such as Google, KaKao, Naver and more. But we began to research using ETRI Toolkit for the following reasons.

- It is known that ETRI toolkit's accuracy is 2~3% higher than Google API in Korean language, and the maximum recognition accuracy reaches 95% in usual conversation. [1]

- ETRI toolkit contains adaptive training functionalities for language and acoustic models with their basic models.[2]

- Cooperation between government institutions and utilizing their R&D results for practical use are being encouraged nationally.

To improve speech recognition system for practical use, it is widely known that the system requires a language model which reflects the grammatical structure for the actual service domain, and the language modeling plays a key role in speech recognition. [3][4][5][6]

A representative language modeling is *n-gram* model, which is the appearance frequency of contiguous sequence of *n* items from a given text corpus based on statistical method.

The toolkit allows users to build their own language model by inputting a text corpus, so our approach was focused on testing the effect of the language model for audio-visual records.

Research Design

Experimental Method

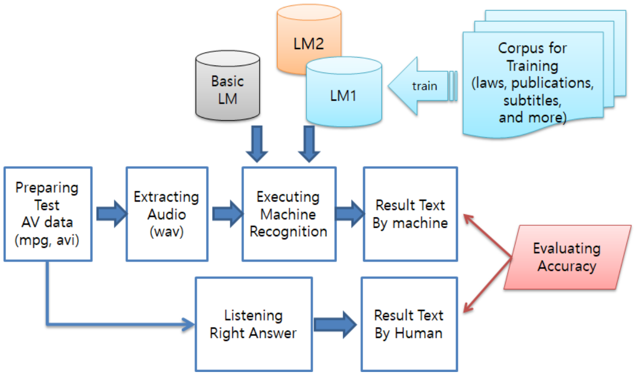

Figure 1*. Research process for our work*

The basic research process is depicted in Figure 1. We tried to compare and analyze the two results between machine generated text and right answer text by humans to calculate the accuracy of recognition. To enhance accuracy, we tried to collect a large text corpus, and trained the speech recognition engine by building several language models. Among various text data within the corpus, we selected the group of text having the greatest relation to the theme of test video records. The newly created language models interpolate the baseline language model containing the text corpus with the general area, which is basically included in the ETRI's speech recognition toolkit.

To achieve best recognition results, we repeated the above test process by changing language models and interpolation weights.

We also performed some acoustic adjustment tests such as audio gain control to find their effects in speech recognition.

Experimental Conditions

Evaluation data was chosen from videos having cultural contents, and the text corpus was collected from the website of National Archives of Korea for language model adaptation.

- Evaluation data: 10 video records created in 1953~1969 having a topic area of culture and economy which were produced for government public relations.
- Training text corpus: 4500 pages of text having topic area of politics, economy, culture and society, which were composed of subtitles of old video records. The size of text file is 13.7Mbytes in ASCII format.

Different interpolation weights were assigned between the new model and original baseline model to test the effect of the new model.

a) [Base LM : New LM] = [0.7 : 0.3]
b) [Base LM : New LM] = [0.3 : 0.7]

We also separated the pages of training data to test the effect of the amount of training data.

- On the assumption that the 4500 pages of training data is the max amount, we built each language model with text pages of 100p, 500p, 1000p, 2000p, 3000p, by extracting segments from original training data.

Experimental Results

Modern News

Before testing on our old video collection, we tested on some modern news broadcasts to get reference values of accuracy and to figure out basic recognition performance of the toolkit.

As shown in table 2, recognition accuracy percentages ranged from 72% to 87%, and the quality of results were generally acceptable to understand the contexts of whole texts even though machine recognition results had several incorrect words.

Table 2. Test results for modern news

No.	Title of article (broadcasting news in 2015 - 2017)	Accuracy in word (%)	Accuracy in syllable(%)
1	Sewol ferry is floating up to the sea	71.76	77.58
2	Abe's position in Japan	72.67	77.19
3	'13m floating up' semisubmersible ship moving	76.4	84.4
4	Sewol ferry's status after three years	76.55	82.45
5	When cars meet a 500 million dollars supercar	75.86	83.14
6	Vote result create a stir in the Democratic primary	87.03	90.45

Language Model Training for Old News

The accuracies of old video records were measured from 23% to 53% with an average of 41.5% only in the case of the baseline language model. The results were considered low due to the quality of old video files which reduced the accuracy of recognition. In terms of acoustic quality, there were constraints such as noise, background music and clipped signals in old video records. There were also limitations in terms of the language model, such as vocabulary not included in the baseline language model and nonstandard language style spoken at that time.

After training the new language model(We named it 'Old news model'), the accuracy was increased to an average of 55.5% as shown in table 3. It is clear that the words and sentences with relevant topic areas have influenced the recognition accuracy.

The results also show the case of having 0.7 of interpolation weight which is slightly better than the case of weight 0.3.

Table 3. Test results for new language model training

No.	Test Records	Accuracy(%) for Baseline Language Model	Accuracy(%) for New Model (Interpolation Weight 0.7)	Accuracy(%) for New Model (Interpolation Weight 0.3)
Accuracy(Average)		**41.487**	**55.504**	**51.871**
1	CEN0002128	22.82	55.48	43.51
2	CEN0002138	35.10	55.36	48.60
3	CEN0002507	41.49	63.29	58.42
4	CEN0002551	46.98	52.92	53.51
5	CEN0002563	53.35	62.88	58.62
6	CEN0002789	47.53	50.52	50.22
7	CEN0003008	50.87	53.39	53.23
8	CEN0003067	36.62	34.58	35.80
9	CEN0003068	39.00	50.16	47.13
10	CEN0003212	41.11	76.46	69.67

Relation between Training Quantity and Accuracy

Collecting the text corpus with relevant topic areas on test data is not simple work, even if the internet is a sea of information. Moreover, it is hard to find text data reflecting historic periods from the internet. We wanted to know how much text data is required to increase recognition accuracy to a useful level. As mentioned in the case of modern news, it seems to be very meaningful that the accuracy reaches to roughly 80%.

Table.4 shows a relation between the quantity of training data and accuracy and the average accuracies manifested in Figure 2.

Table 4. Test results for segmented training data

No.	Test Records	Baseline model (0p)	100p of Old news model	500p of Old news model	1000p of Old news model	2000p of Old news model	3000p of Old news model	4500p of Old news model
Accuracy (Average %)		**41.487**	**42.767**	**43.599**	**48.209**	**51.746**	**53.109**	**55.504**
1	CEN0002128	22.82	25.73	25.62	31.54	54.81	54.47	55.48
2	CEN0002138	35.10	39.06	41.56	43.47	46.99	51.69	55.36
3	CEN0002507	41.49	44.43	44.8	66.79	67.71	66.15	63.29
4	CEN0002551	46.98	48.64	49.81	49.71	52.53	53.8	52.92
5	CEN0002563	53.35	59.03	55.78	59.03	60.24	60.04	62.88
6	CEN0002789	47.53	50.07	47.08	51.12	47.68	49.78	50.52
7	CEN0003008	50.87	47.79	51.1	51.58	52.92	53.71	53.39
8	CEN0003067	36.62	31.98	33.44	33.28	32.63	34.17	34.58
9	CEN0003068	39.00	38.58	39.73	45.05	46.51	48.49	50.16
10	CEN0003212	41.11	42.36	47.07	50.52	55.44	58.79	76.46

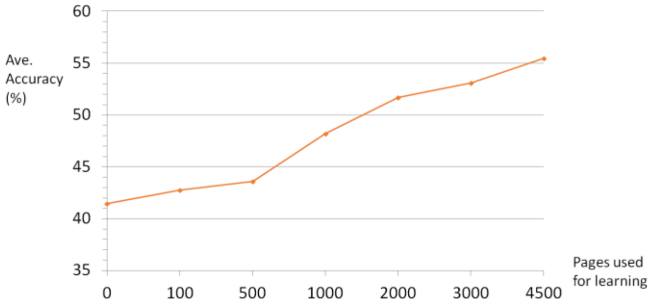

***Figure 2.** A graph of relation between training data and accuracy*

By regression analysis, we made an estimation that the graph of figure 3 is similar to a linear scale. If the assumption is correct, more than 10,000 pages of text are needed to get 80% accuracy.

As such, it is considered that a more relevant corpus is required to increase the accuracy, because the appearance frequency of relevant words will be increased in the n-gram model. But additional research is needed on this matter to find out more precise correlation.

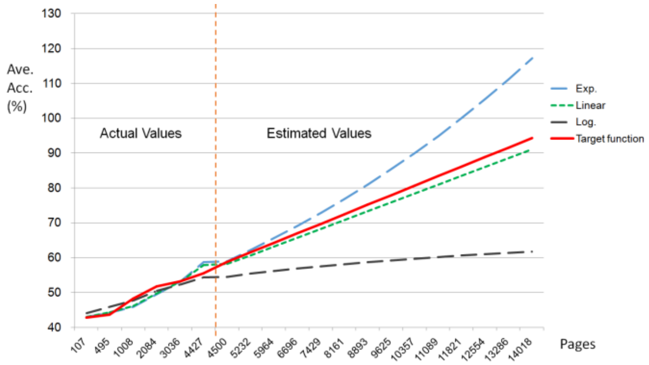

***Figure 3.** Estimation of training quantity*

Narrowing Language Model

While the above-stated test results are related to a relatively large and wide range of text corpus, we tried to narrow the subject area of test records and its training corpus elaborately.

We selected the test video records related to the subject of ‘Records Management’, and gathered training data from laws, regulations, and research reports on records management. The size of training data is 835 pages and 1.2Mbytes, relatively small to the previously mentioned old news model.

Table 5. Test results for the 'Records Mgmt. Model'

No.	Test Records	Baseline Model	‘Old News Model’	‘Records Management’ Model
1	CEQ0000609	52.12	54.36	54.17
2	CEQ0000610	**23.72**	29.77	**31.32**
3	A Promotion Video for the National Archives of Korea	**64.1**	66.95	**69.8**
Average Accuracy(%)		**46.65**	**50.36**	**51.76**

Table 5 shows that there was a 5%p enhancement in the 'Records Management' model rather than the baseline model. The 'Old News' model also brought about a 4%p enhancement than the baseline model, but the 'Records Management' model was slightly better than the 'Old News' model.

It is worth noting that even though the training size of the 'Records Management' model is only 1/10 of the 'Old News' model, the improvement effects are similar between the two models. That means we need to train with the most relevant text corpus in a subject to get the best recognition results. In other words, it is expected that recognition qualities will be better when language models are classified by each subject area, in terms of cost and efficiency.

Acoustic Adjustments

Old video records produced in the 1960s and 1970s sometimes have bad audio signal characteristics. 'Clippings’, exceeded waveforms in input level, are one of the major obstacles for speech recognition.

We adjusted the clipped signal to make it below the maximum limit by using audio editing software as shown in Figure 4.

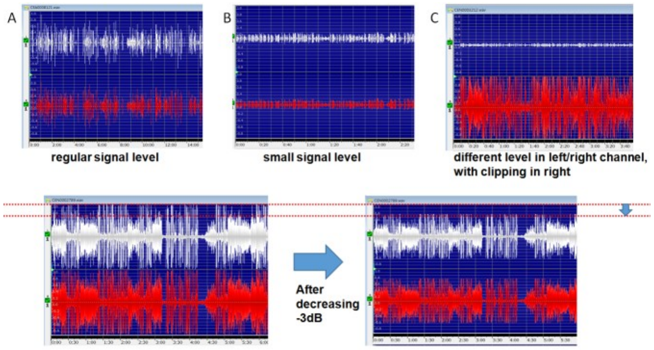

***Figure 4.** Lowering volume gain for clipped signal*

Table 6 shows recognition test results on lowering signal levels, but we can see no meaningful enhancement compared to the unadjusted case. It is considered that signal information could not be expected to be recovered when it was lost already at creation or in conversion stages.

Table 6. Test results for lowering signal levels

No.	**Test Records**	**Acoustic Characteristic**	**0dB (Old News Model)**	**–3dB (Old News Model)**	**–10 dB (Old News Model)**
Accuracy(Average %)			**55.504**	**55.662**	**51.983**
1	CEN0002128	Clipping in right	55.48	51.34	31.88
2	CEN0002138	Normal	55.36	55.36	51.13
3	CEN0002507	Clipping lightly	63.29	63.57	48.76
4	CEN0002551	Normal	52.92	53.80	53.31
5	CEN0002563	Clipping lightly	62.88	63.89	66.53
6	CEN0002789	Clipping heavily	50.52	52.17	53.82
7	CEN0003008	Normal	53.39	53.15	47.48
8	CEN0003067	Clipping heavily	34.58	35.15	35.07
9	CEN0003068	Clipping lightly	50.16	52.14	51.41
10	CEN0003212	Clipping in right	76.46	76.05	80.44

On the contrary to the above case, we amplified the audio signal to be louder for the case of low volume, by using the auto gain function in audio editing software. But there was no enhancement in recognition accuracy after amplification as shown in Figure 5. The reason is considered that noises were amplified together with the normal signal.

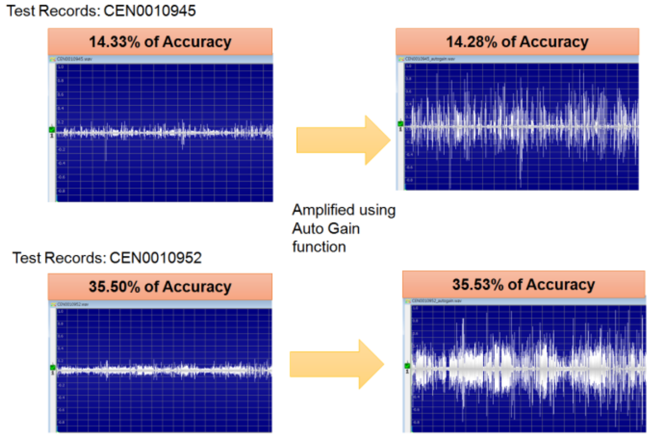

Figure 5. Amplifying gain for low level signal

Design of a Strategy for Applications

Because of technical limits and low quality of audio signals, achieving 100% of recognition accuracy may have a long way to go. We designed a business plan to use speech recognition technology on records management, even though the accuracy is not perfect. When average accuracy reaches more than 60%, we expect that speech recognition technology could be available for business purposes.

We classified service categories according to their applications and users and the development of research to enhance accuracy as shown in table 7.

Table 7. Service categories of speech recognition for audio-visual records management.

Accuracy	For what			For whom		
	Reference materials for Access and Description work	Expanded Indexes for Keyword Searching	Scene Searching	Internal Staffs using Records Mgmt. System	Speech to text service for other institutions	AV records subtitle conversion on website for the public
60~69%	✓	✓		✓		
70~84%	✓	✓	✓	✓	✓	
More than 85%	✓	✓	✓	✓	✓	✓

Obstacles and Further Research for Accuracy Enhancement

In our research to apply speech-to-text technology to old video records, we found various difficult factors on enhancing recognition accuracy for actual use.

Difficulty to collect text corpus for language training

The text material from the internet is limited in type and quantity. Although vocabulary and sentences related to old record contents are required, digitized old newspapers are not helpful because the digitized image could not be converted into text data directly. Also, public records have a great variety of topic areas, such as politics, economics, society, culture, administration, law, history, and more, involving entire government sectors.

Difference between colloquial style and literary style

Video records to be recognized usually contain speeches in a colloquial style such as a public speech, conference, and broadcast script. But the text corpus we could get from the internet is formed in a literary style like a publication. This difference produces some restraints to get best results when using language training.

Various acoustic characteristics in old records

There might have been limitations on broadcasting technology and audio conversion methods several decades ago, which may bring about loss of signal information. Acoustic characteristics are also very different, such as broadcasted videos which include background music, outdoor events, conference rooms, etc.

Comprehensive development of speech recognition technology might overcome these problems someday. But following researches related to building language models are required to develop recognition performance in the application level, as well as, improvement of the speech recognition engine itself.

- Optimum size of text corpus to be collected for language model: How much data do we need to collect considering costs or accuracy?
- Automated and efficient methods to collect text corpus: Are there any methods like web crawling to collect a corpus for a certain subject to deal with any target records?
- Appropriate topic classification for both target records and language models: How to divide topics for each language model such as politics, economy, culture, etc.? And how to designate a certain target record to the corresponding language model?

Conclusions

Speech recognition technologies have been evolving, but old audio-visual records in the 1960s and 1970s still show the difficulty of converting speech to text because of several reasons. To overcome low accuracy of recognition, we used a speech recognizing toolkit based on deep learning and 14%p of accuracy was increased by training relevant corpus data.

To increase recognizing accuracy more and more, there remains various problems to be solved such as collecting a large corpus with related topic areas, overcoming damaged audio signals and interference from background noise.

Despite those problems and obstructions, speech to text conversion work can help records managers and the public to search and understand audio-visual records more easily. We proposed service categories according to each accuracy and the purposes of use.

With continuous researching for recognition accuracy, building a dedicated speech to text conversion system will be the final goal of this research. The system will provide a text conversion service and other useful functions to our archives management system and to the public.

References

[1] Chosun Ilbo(a daily newspaper), "ETRI developed speech to text conversion technology", "http://it.chosun.com/news/article.html?no=2833519", 2017

[2] ETRI's technology transfer website, "Server Based Speech Recognition Technology", http://itec.etri.re.kr

[3] Sang-Hoon Kim, "Speech Recognition Technology, from an Interpreter to a Secretary", Convergence Research Review, vol.3, no.6, pp.5- 34, June 2017

[4] S.H.Na, "Big data for Speech and Language Processing", 2013 Electronics and Telecommunications Trends, pp.52-61, 2013

[5] J.R. Bellegarda, “Statistical LM Adaptation: Review and Perspectives,” Speech Communications., vol. 42, no. 1, pp. 93–108, Jan.2004,

[6] Jeon, Hyung-Bae; Lee, Soo-Young, “Language Model Adaptation Based on Topic Probability of Latent Dirichlet Allocation”, ETRI Journal, vol. 38, no. 3, pp. 487-493, 2016

Author Biography

Jae-Pyeong Kim received his MS degree in Information and Communications Engineering from Chungnam National University, Rep. of Korea, in 2000. From 2004 to 2007, he joined Electronics and Telecommunications Research Institute as a researcher. From 2007, He is working for Archival Preservation and Restoration Center of National Archives of Korea as a researcher with audio-visual archival works.

Yong-Min Shin is working for Archival Preservation and Restoration Center of National Archives of Korea as a researcher.

Sang-Kook Kim is working for Archival Preservation and Restoration Center of National Archives of Korea as a director of electronic and audio-visual archives team.