INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Improving Speech Recognizers by Refining Broadcast Data with Inaccurate Subtitle Timestamps

Jeong-Uk Bang[1], *Mu-Yeol Choi*[2], *Sang-Hun Kim*[2], *Oh-Wook Kwon*[1]

[1]Chungbuk National University, South Korea
[2]Electronics and Telecommunications Research Institute, South Korea
{jubang,owkwon}@cbnu.ac.kr, {mychoi,ksh}@etri.re.kr

Abstract

This paper proposes an automatic method to refine broadcast data collected every week for efficient acoustic model training. For training acoustic models, we use only audio signals, subtitle texts, and subtitle timestamps accompanied by recorded broadcast programs. However, the subtitle timestamps are often inaccurate due to inherent characteristics of closed captioning. In the proposed method, we remove subtitle texts with low subtitle quality index, concatenate adjacent subtitle texts into a merged subtitle text, and correct the timestamp of the merged subtitle text by adding a margin. Then, a speech recognizer is used to obtain a hypothesis text from the speech segment corresponding to the merged subtitle text. Finally, the refined speech segments to be used for acoustic model training, are generated by selecting the sub-parts of the merged subtitle text that matches the hypothesis text. It is shown that the acoustic models trained by using refined broadcast data give significantly higher speech recognition accuracy than those trained by using raw broadcast data. Consequently, the proposed method can efficiently refine a large amount of broadcast data with inaccurate timestamps taking about half of the time, compared with the previous approaches.

Index Terms: data refinement, data selection, text-to-speech alignment, speech recognition

1. Introduction

As the deep learning paradigm is applied to speech recognition systems, database reinforcement has a greater impact on performance than algorithm development. However, most speech databases are built manually, which consumes a lot of money and manpower. For this reason, major research institutes and global companies are working on building databases automatically, and in recent years, efforts have been made to build a more spontaneous speech databases [1, 2, 3].

Broadcast data contain a great deal of spontaneous speech data useful for training speech recognition systems. Broadcast data can be easily used in data refinement experiments, since they contain transcripts for the hearing impaired and metadata such as metadata such as speaker changes, timestamps, music and sound effects, TV genre of each show according to the method of collection [1].

In the recent works described in [1, 4], a large amount of broadcast data with various metadata are collected in collaboration with broadcasting stations in order to improve the performance of speech recognition systems. Here, the various metadata are useful to detect speech segments for each speaker. To extract correct timestamps of the detected segments, a speaker-adaptive decoder is used to obtain a hypothesis text of the each speech segment, and then the decoder output is compared with original subtitle text to identify matching sequences. Non-matching word sequences from the original subtitle text are force-aligned to the remaining speech segments. Finally, refined speech segments are generated by selecting the appropriate speech segments for acoustic model training.

In this work, broadcast data consists of only audio signals, subtitle texts, and their timestamps, because we extract broadcasted programs directly on air. Hence, the subtitle timestamps are often inaccurate due to inherent characteristics of closed captioning. To apply the previous methods [1, 4] to our broadcast data at hand, voice activity detection and speaker diarization toolkits are required. However, even though the toolkits are developed elaborately, it is difficult to expect good performance for our broadcast data, since the broadcast data consist of multi-genre programs, contain various noise and music, and are lacking in the metadata such as the number of speakers, music and sound effects [5]. For this reason, we aim to efficiently extract speech segments that are useful for training acoustic models in the case that only inaccurate subtitle timestamps are given, for the purpose of processing multi-genre broadcast data collected every week. Our method differs from the previous methods in that we do not use metadata or auxiliary toolkits to obtain the correct timestamp of all speech segments.

In Section 2 we describe the broadcast data used for our experiments. The details of the proposed method are explained in Section 3, and experimental results are shown in Section 4. Finally we draw conclusions in Section 5.

2. Broadcast Data

The broadcast data used in this work consist of multi-genre recordings of 3,137 hours broadcast on 7 major broadcasting channels of South Korea from March to June 2016. The broadcast data were recorded in a program unit, but often included advertisements or other programs without subtitles in the front or back part of audio signals. Furthermore, some parts of audio signals did not have subtitles in case of interviews, lyrics, or sports broadcasting. In order to evaluate the performance of acoustic models, we used manually-segmented evaluation data consisting of five genres: News, culture, drama, children (Child.), and entertainment (Ent.). The evaluation data were not included in the training set for speech recognition.

The audio length and number of subtitles for each genre in the evaluation data set are shown in Table 1. The raw audio signals have a length of about an hour for each genre, and audio parts without subtitles appeared in advertisements,

Copyright © 2017 ISCA

http://dx.doi.org/10.21437/Interspeech.2017-650

songs, long silences, and so on. The ‘Filtered’ data mean the portion where speech actually existed. From the table, we know that 43% of raw broadcast data can be refined on average.

We removed short subtitle texts with a duration of less than one second. As a result, whereas most of subtitles in the news and culture genres were unchanged, many subtitles in the entertainment genre were moved. This means that the data in the entertainment genre have a lot of short utterances, and accordingly are harder to extract actual timestamps and recognize correctly than those in the other genres.

To check the quality of subtitles, we first compare the beginning and ending times of the subtitles with those of the actual corresponding speech. We observed that the beginning and ending times of the actual speech were about 6 seconds and 8 seconds earlier that the subtitle time, respectively. In terms of transcripts, there were about 2% of transcription errors due to incorrect translation of foreign utterances and wrong transposition of words.

Table 1: *Audio size and number of subtitles in the evaluation data set.*

Genre	Audio size (hh:mm)		Number of subtitles	
	Raw	Filtered	Raw	Filtered
News	01:08	00:40 (58%)	412	404 (98%)
Culture	01:14	00:30 (41%)	450	408 (91%)
Drama	01:04	00:20 (31%)	684	430 (63%)
Child.	00:43	00:14 (33%)	446	263 (59%)
Ent.	01:24	00:40 (47%)	1,744	857 (49%)
Average	01:07	00:29 (43%)	747	472 (63%)

3. Proposed Method

As shown in Figure 1, the proposed method consists of five steps: Text normalization, speech segment extraction, speech recognition, text alignment, and data selection.

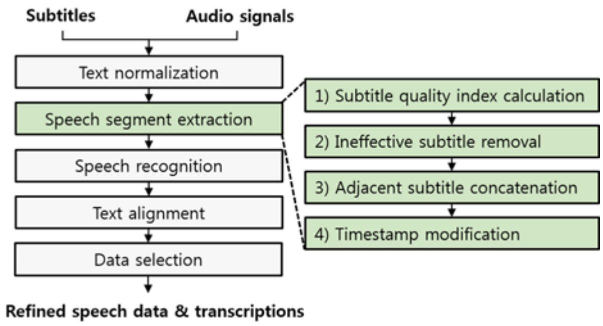
Subtitles
Audio signals
Text normalization
Speech segment extraction
Speech recognition
Text alignment
Data selection
1) Subtitle quality index calculation
2) Ineffective subtitle removal
3) Adjacent subtitle concatenation
4) Timestamp modification
Refined speech data & transcriptions

Figure 1: *Block diagram of proposed method.*

3.1. Text normalization

In the text normalization step, subtitles containing invalid characters or lots of English words are removed, and then the subtitle texts are converted into a morpheme sequence so that they can be aligned with Korean morpheme-based speech recognition outputs.

3.2. Speech segment extraction

In this step, we extract the most likely speech interval from the input audio. The step is divided into 5 sub-steps: Subtitle quality index (SQI) calculation, ineffective subtitle removal, adjacent subtitle concatenation, and timestamp modification of the merged subtitle text. Selecting appropriate speech segments is critical in reducing the decoding time of speech recognizers that should process a bulky amount of data.

Since broadcast data can be easily obtained at any time, it is better to remove the speech segments with lower quality rather than refining. For this reason, the subtitle quality index (SQI) of a subtitle text is defined as the ratio between the duration and the number of characters in a subtitle text

$$SQI = \frac{duration\ of\ subtitle\ timestamp}{number\ of\ characters\ in\ subtitle\ text} \quad (1)$$

This index indicates the duration of audio signals required to refine a character in a subtitle. A subtitle text with long duration but with few characters shows an extraordinary high value. Thus, we remove the subtitles having SQI larger than 1 considering the average speech rate and the time-lag of subtitles.

Some examples of the removed subtitles are shown in Table 2. The subtitle text “Thank you for watching” with extremely long duration usually occurs when the subtitle timestamp has an ending-time error at program endings. In this case, it takes a lot of time to find the corrected timestamp of actual speech within the given duration of 594 seconds. This kind of abnormally long subtitles are often found at program endings, scene changes, or dialog disconnection.

Table 2: *Examples of removed subtitles.*

Subtitles (Translated into English)	Duration (s)	SQI
Thank you for watching	594	31.3
We will see you next week	215	10.8
Who is it	10	1.4
Okay	37	9.3

Whereas correct subtitles include speech segments corresponding to the text within the given timestamp, incorrect subtitles may not contain actual speech within the given timestamp. For this reason, we concatenate adjacent subtitles to prevent search ranges from overlapping. Then, we add margins in front and back of each speech segment corresponding to the concatenated subtitle: -6 seconds to the beginning time and +2 seconds to the ending time of the timestamp. We call this a ‘modified speech segment’.

Figure 2 is an example of the speech segment extraction step, where “s_n” is the n-th subtitle and “m_n” is the n-th preprocessed subtitle corresponding to the modified speech segment. Here, the darker is the color of each segment, the smaller is the SQI value of that segment.

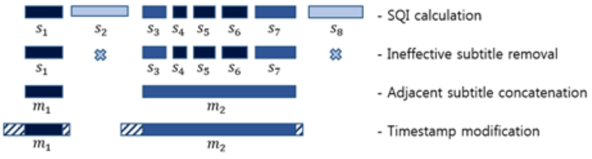

Figure 2: *Example of speech segment extraction.*

3.3. Speech recognition

In the speech recognition step, a speech recognizer produces the word sequence and its timestamp from the modified speech segments. The speech recognizer uses a biased language model (LM) [1], and a vanilla deep neural network (DNN)-based acoustic model (AM). The AM was trained by using 925 hours of manually transcribed speech data in the travel domain. The biased LM is generated at each moment from the sentence obtained by merging two subtitle texts in each program in order to correctly compute the beginning and ending probabilities of each subtitle text. The vocabulary is chosen to include all words occurring in the original subtitle texts. The speech recognizer outputs the time information for each word.

In the previous study [6], speaker diarization was applied to the entire input audio stream, and then a speaker-adaptive speech recognizer in two-pass recognition framework was used to decode. On the other hand, in our experiments we only perform one-pass speaker-independent speech recognition.

3.4. Text alignment

In the text alignment step, the preprocessed transcript and the speech recognition output (hypothesis) are aligned. The hypothesis in our experiments usually have more words than the transcript due to the time margin added. In this case, it is common to use a local alignment algorithm that finds best substrings in one sequence that aligns well with a substring in the other [7]. However, the local alignment methods based on local similarity were not appropriate because the similar word sequences frequently appear in a broadcast program.

In the proposed method, we first search for the longest common subsequence (LCS), and recursively aligns the left and right word sequences to find the next LCS. If no LCS does exist, the word sequences are force-aligned to the remaining sequence using the Needleman-Wansch (NW) algorithm [8]. Our method has approximately 2% more detection than the Smith-Waterman algorithm [7] by changing the local alignment to a global alignment problem, because it holds reliable reference positions by the detected LCS.

The proposed method for alignment works as shown in Figure 3. First, the LCS (C_1) is searched between transcripts and hypothesis, and then the alignment table is divided into 3 sub-tables: Left (L_1), right (R_1), and center (C_1). Next, we search for the next LCS in the L_1 and R_1 tables successively. If the LCS exists, the corresponding table is divided into the left and right tables again as shown in L_3 and R_3. Otherwise, a similar string is aligned using the NW algorithm as shown in the L_1 table. These procedure are repeated recursively.

3.5. Data selection

Speech data suitable for acoustic model training are selected when the transcripts and the corresponding speech signals are perfectly matched. Thus, we first split the modified speech segment in the original subtitle unit, and then extract the corrected speech segments beginning from the first matching words and ending at the last matching words corresponding to the timestamp of the hypothesis within the original subtitle text boundary. Here, a few words at the beginning and ending parts of the original subtitle text may be removed from the corrected speech segment. Accordingly, the corrected speech segment may contain some transcription errors because they are generated based on the words between the first matching subsequence and the last matching subsequence.

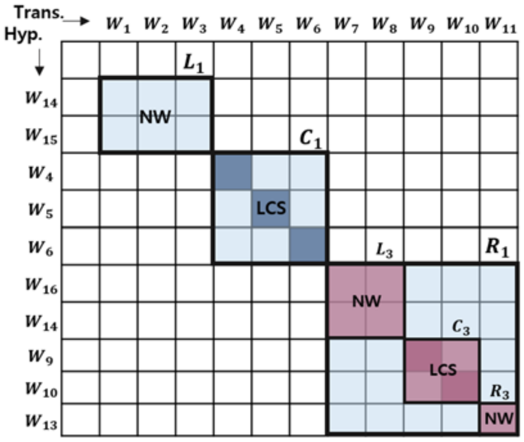

Figure 3: *An example of text alignment.*

In this paper, we select only the word sequence of corrected transcripts without transcription errors by checking the ratio of the number of matched words over the total number of words. The finally selected data are called 'refined speech segments', which will be used for training new acoustic models.

4. Experimental Results

4.1. Experimental setup

All recognition experiments were conducted using the Kaldi toolkit [9]. The input features are equivalent to un-adapted, un-normalized 40-dimensional log-mel filterbank features, spliced for ±7 frames. The acoustic models used are DNN trained by layer-wise back-propagation supervised with a 3-state left-to-right hidden Markov model (HMM). The DNNs use a vanilla system with a learning rate of 0.001 for the input layer with 15*40 nodes, 7 hidden layers using tanh activation function, the output layer with 8,033 nodes using the softmax activation function.

The language model was trained with Kneser-Ney discounting (cut-off 0-3-3) using the SRILM Toolkit [10], and included 1M unigrams, 16M bigrams, and 12M trigrams from broadcast subtitles excluding the evaluation data set. The speech decoder was set to have an acoustic model weight of 0.077, a beam size of 10.0, and a lattice beam of 5.0. The recognition results were compared by calculating the word error rate (WER) in the morpheme unit.

4.2. Experimental results

The broadcast data (raw data) of 3,137 hours were refined based on three methods. In the first method ('TS'), we extracted the modified speech segments (modified data) using the subtitle timestamps without margin processing. In the second method ('TS+MG'), we added margins at the front and back of each subtitle. In the third method ('Proposed'), the modified speech segments were extracted according to the proposed method.

The total length of Korean (KOR) speech segments refined by each method is shown in Table 3. In case of using only subtitle timestamps, only 360 hours of speech segments could be extracted into refined speech segments (refined data), because there is no actually speech in the modified speech segments due to inaccurate subtitle timestamps. In the case where the experiment is performed by adding a margin to increase the possibility that actual speech exists in the speech segments, it produced 903 hours of refined speech segments. But it takes a lot of time to refine it, because the size of the modified speech segments are too large.

The proposed method produces relatively few modified speech segments for 2,383 hours and show the refining result of 939 hours. This produces more refined speech segments with less modified speech segments than just adding margins to timestamps. This means that unnecessary audio signals for refinement have been removed, and audio signals that have subtitle texts but are out of the modified speech segment despite the addition of margins are complemented by connecting subtitles.

Table 3: *Data length (h) after each step (KOR).*

	NR	TS	TS+MG	Proposed
Raw data	3,137	3,137	3,137	3,137
Modified data	2,119	2,119	5,367	2,683
Refined data	2,119	360	903	939

The performance of the evaluation data were compared using the acoustic models trained with each refined speech segments. The evaluation data set was composed of news, culture, drama, children, and entertainment genres as shown in Table 1. To confirm the performance of the non-refined speech segments ('NR'), the acoustic model was trained by segments of 2,119 hours which are previously used as the modified speech segments of the 'TS' method where only the ineffective subtitles were removed in the speech segment extraction sub-steps. As a result, recognition performance was not improved even though we used a large amount of speech data, because the incorrect subtitles often do not contain actual speech within the given timestamp.

Table 4: *WER (%) for Korean language.*

	NR	TS	TS+MG	SUP	Proposed
News	72.2	14.5	13.8	20.7	13.6
Culture	87.4	51.3	48.3	63.4	47.3
Drama	86.7	43.7	40.2	53.8	40.1
Child.	73.0	38.0	35.9	49.8	35.1
Ent.	95.3	75.4	72.2	85.8	71.1
Average	81.2	38.7	36.6	48.0	36.0

Table 4 shows the WER for Korean language. The 'TS', 'TS+MG', and 'Proposed' methods reduced recognition word error rate (WER) to 38.7%, 36.6%, and 36.0%, respectively. The results for the news genre show much better performance than the other genres, because speech data in the news genre mostly read speech with a long duration and small noise. On the contrary, the entertainment genre data show poor performance because the data are mostly uttered in the spontaneous manner and have a short duration, a lot of noise such as background music.

In the matched-pairs sentence segment word-error test using the NIST Scoring Toolkit (SCTK), the proposed method yielded a p-value less than 0.05 in comparison with the 'TS+MG' method and a p-value less than 0.001 in comparison with the other methods. This justifies the statistical significance of the proposed method.

The length of modified speech segment to be processed is proportional to the processing time of the refining system. In the experiment with the margin added to the timestamp, the processing time is the largest. However, in the experiment using only the timestamp, a small amount of broadcast data is refined. Although the proposed method yielded modest improvement of recognition accuracy, it successfully refined a large amount of broadcast data with inaccurate subtitle timestamps taking about half of the time compared with the previous methods. Therefore, it is useful for broadcasting data processing where bulk speech data can be collected every hour.

4.3. Performance comparison with supervised AM

To confirm the performance of the proposed method, further recognition experiments were conducted using a supervised acoustic model (SUP), which were trained by using 925 hours of the manually transcribed speech data. This model is the same as the acoustic model used in the speech recognition step of refining experiment. The experiment using SUP database showed an average WER of 48.0%, and the performance for each genre was similar to the acoustic models trained without using any manual transcription.

Whereas the size of 'SUP' database (925 hours) is similar to the refined data size (939 hours) obtained by the 'Proposed' method, the 'Proposed' method reduced the WER from 48.0% to 36.0%, which is relative error rate reduction of 25.0%. We note that the performance of the 'SUP' and 'Proposed' methods cannot be exactly compared because the training data are different in both methods.

5. Conclusions

This paper focused on efficient refinement of broadcast data with inaccurate subtitle timestamps. The proposed method significantly improved speech recognition performance compared with non-refined speech segments. Compared with the previous method, a large amount of broadcast data having inaccurate subtitle timestamps were efficiently refined in about half of the time. The proposed method can be applied to speech recognition systems that have to be updated frequently because refined speech segments are efficiently extracted from broadcast data. For further study, we plan to confirm the validity of the proposed method for other languages and investigate the performance of unsupervised training methods without any subtitle texts or timestamps.

6. Acknowledgements

This work was supported by Electronics and Telecommunications Research Institute (ETRI) grant funded by Korea government [Strengthening competitiveness of automatic translation industry for realizing language barrier-free Korea]. This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No.2015R1D1A3A01020817).

7. References

[1] P. Lanchantin, M.J.F. Gales, P. Karanasou, X. Liu, Y. Qian, L. Wang, P.C. Woodland and C. Zhang, "The development of the Cambridge University alignment systems for the Multi-Genre Broadcast challenge," in *Proc. Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 647-653.

[2] O. Kapralova, J. Alex, E. Weinstein, P. Moreno, and O. Siohan, "A big data approach to acoustic model training corpus selection," in *Proc. INTERSPEECH,* 2014, pp. 2083-2087.

[3] H. Liao, E. McDermott, and A. Senior, "Large scale deep neural network acoustic modeling with semi-supervised training data for YouTube video transcription," in *Proc. Automatic Speech Recognition and Understanding (ASRU)*, 2013, pp. 368-373.

[4] P. Lanchantin, M.J.F. Gales, P. Karanasou, X. Liu, Y. Qian, L. Wang, P.C. Woodland and C. Zhang, "Selection of multi-genre broadcast data for the training of automatic speech recognition systems," in *Proc. INTERSPEECH,* 2016, pp. 3057-3061.

[5] X. Bost, G. Linares, and S. Gueye, "Audiovisual speaker diarization of TV series," in *Proc. Acoustics, Speech and Signal Processing (ICASSP),* 2015, pp. 4799-4803.

[6] P. Bell, M.J.F. Gales, T. Hain, J. Kilgour, P. Lanchantin, X. Liu, A. McParland, S. Renals, O. Saz, M. Wester and P.C. Woodland, "The MGB Challenge: evaluating multigenre broadcast media transcription," in *Proc. Automatic Speech Recognition and Understanding (ASRU)*, 2015, pp. 687-693.

[7] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *Journal of molecular biology*, vol. 147, no. 1, pp. 195-197, 1981.

[8] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of molecular biology,* vol. 48, no. 3, pp. 443-453, 1970.

[9] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, and K. Vesely, "The Kaldi speech recognition toolkit," in *Proc. Automatic Speech Recognition and Understanding (ASRU)*, 2011.

[10] A. Stolcke, "SRILM – An extensible language modeling toolkit," in *Proc. INTERSPEECH*, 2002.