Automatic Closed-Caption Alignment Using Pronunciation of Speech Recognition Transcripts for Public Relations TV Program

Shinya Takahashi, Tsuyoshi Morimoto * and Yoshiyuki Nishimoto †

Abstract— **When closed-caption texts attach to a video data, it is important to decide the timing of the closed-caption. For this problem, speech recognition results are generally used to match the closed-caption texts with speeches in the video. However, it is not easy to detect matching points for spontaneous speeches due to a lot of redundant words and restatements. Moreover, the speech with background noises and indistinct pronunciations makes it so difficult to detect the voice activity and gives a serious effect on mis-recognition. This paper proposes a method for calculating the correspondence between a phoneme sequence of the closed-caption text and a phoneme sequence of the recognition result using the confusion matrix of the recognized phonemes. To show the effectiveness of this approach, some experiments are demonstrated for public relations TV programs. The experimental result shows that the proposed method is effective to decide the timing of the closed-caption in such case the closed-caption reflects for speech contents sufficiently.**

Keywords: Closed-caption alignment, Confusion matrix, DP matching, Speech recognition, TV program

1 Introduction

Recently, the development and spreading of closed-caption on TV program are the important issues for hearing-impaired people to access more information easily, so the Japan government (the Ministry of Internal Affairs and Communications) amended the Broadcast Law to expand the closed-captioning programs. Now, 40.8% of NHK (Nippon Hoso Kyokai or Japan Broadcasting Corporation) and 27.5% of the main 5 commercial TV stations have the closed-captioning programs[1]. Moreover, terrestrial digital broadcasting that has started from 2003 in Japan is providing some additional information with TV programs as a standard function, so the closed-caption and other information are beginning to be utilized widely. Besides, a framework that can provide the closed-caption texts to streaming videos on Web browsers easily was standardized, so the closed-caption service on Web site is expected to spread more and more (Fig.1).

There are the following two problems when attaching the closed-caption texts to the video data[2], [3]: (1) what content is each closed-caption text given? (2) which timing is each closed-caption text displayed at? In addition, the former problem requires to modify each text into some appropriate form and length after dictation by hand or a speech recognizer in the case no manuscript text exists. On the other hand, for the latter problem, there are quite a few case that transcripts or manuscript texts can be utilized, so it makes important to calculate the correspondence between the video and the closed-caption text automatically and accurately.

Figure 1: Example of closed-caption.

To cope with this correspondence problem the method matching the closed-caption texts with the result of speech recognition is used generally. For example, a method for matching the words in the closed-caption texts with that in the speech recognition results has been proposed in [4], [5]. As another example in [6], they have

*Dept. EECS, Fukuoka University, 8-19-1 Nanakuma, Jonan-ku, Fukuoka-shi, 814-0180 Japan (Tel: +81-92-871-6631(ext.6572); Fax: +81-92-871-6031; Email: takahasi@tl.fukuoka-u.ac.jp).
†J-FIT Co. Ltd., Fukuoka, Japan.

proposed a method for matching the length of the closed-caption with that of the speech recognition results.

In these approaches, however, it is not easy to detect matching points between spontaneous speeches and the closed-caption text due to a lot of redundant words and restatements. These are included in the spontaneous speeches but are removed from the closed-caption texts. In addition, there is another problem that background noises and/or indistinct pronunciations make it difficult to detect the voice activity and to perform speech recognition accurately.

In order to cope with these issues, in this paper, we propose a method for calculating the correspondence of a phoneme sequence of the closed-caption texts and a phoneme sequence of the recognition results using a confusion matrix and we conduct some experiments for evaluation of performance.

2 Automatic Closed-Caption Alignment

A basic idea is to detect appropriate points corresponding to the start point of the closed-caption text from sound waves using DP (Dynamic Programming) matching.

The detail of the procedure is as follows.

1. derive sound data from video data.
2. divide the sound into some waves with an appropriate length.
3. recognize each sound with a speech recognizer.
4. obtain a phoneme sequence with corresponding time from all results of speech recognition.
5. convert a closed-caption text into a phoneme sequences of pronunciation.
6. calculate a correspondence of the phoneme sequences of the recognition results and that of the closed-caption text with DP matching algorithm.

Here, in the second process, the partition of the sound waves is performed by comparing a threshold of the power-difference (70dB is used here) [1].

Note that the purpose of this partition process is for the constraint of the speech recognition, not for the voice activity detection. A method for calculating the correspondence of each utterance after detecting the voice activity has also been proposed. However, the accuracy of the voice activity detection strongly depends on background noises and a recording environment. So, in this paper the correspondence is calculated without the voice activity detection.

In the third process, the speech recognition is conducted with unigram language model only. The reason of not using bigram and trigram language models is that a lot of substitution errors for words tend to occur due to weights for these language models.

3 Algorithm for Matching

3.1 DP matching

Let $\mathbf{A} = \{a_i | (i = 0, \ldots, N)\}$ and $\mathbf{B} = \{b_j (j = 0, \ldots, M)\}$ denote the phoneme sequence of the closed-caption text and the phoneme sequence obtained from the recognition result, respectively.

The optimal distance between $\mathbf{A}$ and $\mathbf{B}$ can be obtained by calculating the following DP recursive equation,

$$D(i,j) = \min \begin{cases} D(i,j-1) + p_{\text{ins}}, \\ D(i-1,j-1) + d(i,j), \\ D(i-1,j) + p_{\text{del}}. \end{cases} \quad (1)$$

Here, $d(i,j)$ is a local distance between phonemes a_i and b_j defined by using a confusion probability $p(b_j|a_i)$ as follows:

$$d(i,j) = 1 - p(b_j|a_i).$$

As the penalties of insertion and deletion $p_{\text{ins}}, p_{\text{del}}$ for vowels and consonants, we used $p^{(\text{v})}_{\text{ins}}, p^{(\text{c})}_{\text{ins}}, p^{(\text{v})}_{\text{del}}, p^{(\text{c})}_{\text{del}}$, because it can be assumed that the occurrence probabilities of them are differ[2].

3.2 Confusion Matrix

As described above, it is so difficult to recognize low quality speech signals with a lot of noises or background sounds, so the correct correspondence of the phonemes can not be calculated accurately. To cope with this problem, we attempt to utilize a confusion matrix for calculating the local distance in Eq.(1). Each element of the confusion matrix is calculated as a probability that each phoneme is recognized correctly or is misrecognized as other phonemes, which is obtained from training speech data in advance. For training data we used the speech data of lectures included in Corpus of Spontaneous Japanese[7]. To be concrete, we used 74 minutes data that consist of the following six lecture data: A01F0055, A01F0067, A01F0122, A01F0132, A01F0143, A01F0145.

Phonemes used in experiments are shown in Table 2. We calculate the confusion probabilities between the 41

[1] To calculate a power of speech signals, 150ms windows size and 10 ms frame length are used.

[2] (v) and (c) mean vowels and consonants, respectively.

Table 1: Data used in experiments

Data	broadcasting date	broadcasting time	# of CC sentence	# of word in CC	# of word in speech	word correct rate of speech recognition
A	6/08*	8min.19sec.	49	1284	1429	26.7%
B	6/16**	3min.	19	489	562	32.2%
C	7/21**	3min.	22	536	585	21.0%
D	8/3*	8min.12sec.	49	1175	1281	28.0%

* Komyu! Fukuoka ** Gimon Kaiketsu! Fukuoka Q

Table 2: Japanese phonemes used in experiments

vowels	/a/,/e/,/i/,/o/,/u/
long vowels	/a:/, /e:/, /i:/, /o:/, /u:/
consonants	/b/, /by/, /ch/, /d/, /dy/, /f/, /g/, /gy/, /h/, /hy/, /j/, /k/, /ky/, /m/, /my/, /n/, /ny/, /p/, /py/, /r/, /ry/, /s/, /sh/, /t/, /ts/, /w/, /y/, /z/
moraic silence	/Q/
moraic nasal	/N/

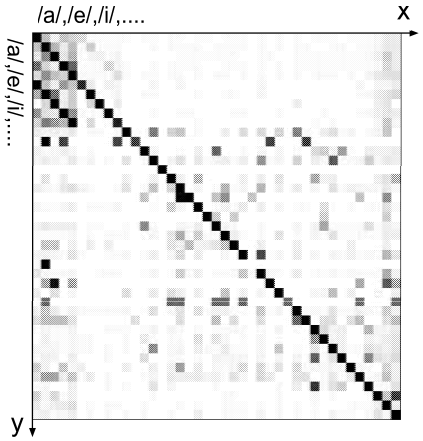

Figure 2: Calculated Confusion Matrix

phonemes that include the phonemes shown in Table 2 and a short-pause for the confusion matrix.

Fig. 2 shows the calculated confusion matrix. Each pixel in the matrix represents a probability of $p(x|y)$. The matrix is arranged in order of the phonemes shown in Table 2 from the top-left corner to the bottom and the right. The darker pixel indicates the higher probability. Note that the contrast of the image in this figure is enhanced for visibility.

As this figure shows, vowels tend to confuse other vowels, especially long vowels, rather than consonants. Besides, comparatively many confusions between plosives like /b/, /d/, /p/ can be seen from this figure.

4 Experiments

4.1 Experimental Condition

In the experiments, we used a Japanese large vocabulary speech recognizer Julius[8] ver. 3.4, a language model trained from Web documents with 60,000 words and an acoustic model with a gender independent HMM in Julius dictation kit Ver.3.1. Julius is one of the most famous speech recognizer and is used for not only Japanese but also English, French, Slovenian, Thai language, and so on. These recognizer and models can be downloaded in Julius Web site[9].

Table 1 shows details of 4 TV programs used in the experiments. These programs were broadcast in 2007 and are opened to the public at official Web site[10]. Here, "# of word in CC" means the total number of words included in the closed-caption texts and "# of word in speech" means the total number of words included in speech data. The reason of that "# of word in CC" is about ten percent less than "# of word in speech" is because redundant words were removed when the closed-caption texts have been created.

Table 1 also shows word correct rates in speech recognition. The word correct rates are the result of the speech recognition experiments with the trigram language model and are calculated with transcriptions written by hand as correct sentences. Note that the speech recognition experiments here are conducted without removing noise and non-speech parts using a trigram language model.

As shown in the table, the word correct rate of speech recognition resulted in very low. The reason of this result is that the target speech included not only a lot of background noises and overlapped sounds but also a lot of redundant words and restatements peculiar to spontaneous speech.

Examples of a transcription and a closed-caption text are shown in Table 3. In this table, the words removed from the closed-caption texts, such as redundant words and restatements, are shown underlined and corrected words are shown in the bold style. As can be seen from this example, in the transcripts of the speech have a lot of the redundant words.

Table 3: Example of transcription and closed-caption text.

Transcript	なるべく、えっと、デパートとか、あの、販売店の紙袋とかね、ビニール袋も、なるべく、もう、買物袋にまとめてね、もらわないようにはしてます。 (Well, I bring my own shopping bag, if possible, instead of using a paper and ah well, a plastic bag of department store, ah, and some shops if possible.)
CC text	デパートとか、販売店の紙袋やビニール袋は、なるべく、買物袋にまとめて、もらわないようにしてます。 (I bring my own shopping bag instead of using a paper and a plastic bag of department store and some shops if possible.)

Table 4: Result of phoneme recognition

Data	# of recognized phonemes	# of phonemes in CC	ratio of number
A	6099	5274	1.16
B	2370	2040	1.16
C	2309	2111	1.09
D	5688	4516	1.26

4.2 Experimental Results

Table 4 shows the result of phoneme recognition.

Using the DP algorithm described in section 3.1, we calculated the correspondence between the recognized phoneme sequence and the phoneme sequence of the closed-caption text. For comparison, we conducted the following 4 experiments:

(1) without confusion matrix and with fixed parameters ($p_{\mathrm{ins}}^{(\mathrm{v})} = p_{\mathrm{ins}}^{(\mathrm{c})} = p_{\mathrm{del}}^{(\mathrm{v})} = p_{\mathrm{del}}^{(\mathrm{c})} = 1.0$)

(1') without confusion matrix and with the best-scoring parameters ($p_{\mathrm{ins}}^{(\mathrm{v})} = 0.75, p_{\mathrm{ins}}^{(\mathrm{c})} = 0.75, p_{\mathrm{del}}^{(\mathrm{v})} = 1.0, p_{\mathrm{del}}^{(\mathrm{c})} = 0.5$)

(2) with confusion matrix and fixed parameters ($p_{\mathrm{ins}}^{(\mathrm{v})} = p_{\mathrm{ins}}^{(\mathrm{c})} = p_{\mathrm{del}}^{(\mathrm{v})} = p_{\mathrm{del}}^{(\mathrm{c})} = 1.0$)

(2') with confusion matrix and the best-scoring parameters ($p_{\mathrm{ins}}^{(\mathrm{v})} = 0.75, p_{\mathrm{ins}}^{(\mathrm{c})} = 0.75, p_{\mathrm{del}}^{(\mathrm{v})} = 0.75, p_{\mathrm{del}}^{(\mathrm{c})} = 0.5$)

Table 5: Results for alignment accuracy

Data	Method (1)	(1')	(2)	(2')
A	0.556	0.502	0.493	0.475
B	0.305	0.318	0.273	0.259
C	0.838	0.422	0.827	0.532
D	1.035	1.000	1.399	1.400
Average	0.684	0.561	0.748	0.546

Values of each parameter in (1') and (2') were decided as those in the best of the average score for all 4 data where the parameters were given from 0.25 to 1.0 with 0.25 step.

The following equation is used as the local distance between phonemes for the above (1) and (1') cases.

$$d(i,j) = \begin{cases} 0 & \text{if } a_i = b_j, \\ 1 & \text{otherwise.} \end{cases}$$

Table 5 shows the mean error values between the start time obtained by the proposed method and that given by hand.

As can be seen from this table, comparatively good results were obtained for data A,B, and C by using the confusion matrix. On the other hand, the result for data D did not indicate the sufficient effectiveness of the confusion matrix. As shown in Table 4, the reason is supposed to be that the ratio of the number of the recognized phonemes and that of the phonemes in the closed-caption text is ten percent more than the ratio of the other data because the phonemes of the recognition result include a lot of unnecessary phonemes not corresponding to that of the closed-caption.

Finally, Table 6 shows examples of the closed-caption alignment.

5 Conclusions

In this paper, we proposed the method that can obtain the alignment between a phoneme sequence in the closed-caption and a recognized phoneme sequence using the confusion matrix. The experimental result showed that the proposed method is effective for phoneme alignment in such case the closed-caption reflects for speech contents sufficiently.

In this paper, we did not evaluate the timing that the closed-caption should be removed. To detect the segment that the closed-caption should not be displayed, it is need to distinguish the necessary phonemes and the unnecessary ones. It is supposed to be difficult to realize the detection of the unnecessary segments using the approach with the recognition result only. In the future, we

Table 6: Examples of closed-caption alignment.

CC text	m i n a s a N - m o - sp 皆さんも (all of you (viewer))
Rec. result	n i r a s a: N u m o a sp ニラさぁんうモア (N/A)
CC text	sh i t e m i m a s e N - k a してみませんか (Do you try . . . ?)
Rec. result	s - - e N - n a s a N sp t a せんなさんた (N/A)

intend to investigate a method for calculating the timing of removing the closed-caption with the length of the closed-caption text or the number of moras in the closed-caption text. In addition, we also intend to utilize the scene change information obtained from image processing to calculate the display timing of the closed-caption accurately.

Acknowledgment

We would like to thank the public relations section of Mayer's office in Fukuoka city for their permission for us to use TV program data. This research was partially supported by the Ministry of Education, Science, Sports and Culture, Grant-in-Aid for Young Scientists (B), 19700184, 2007.

References

[1] the Ministry of Internal Affairs and Communications, "A reports on the research of digital broadcasting for visually or hearing impaired people (in Japanese)," http://www.soumu.go.jp/s-news/2007/pdf/070330_19_ts2.pdf

[2] K. Watanabe, M. Sugiyama, "Time Alignment between Caption and Acoustic Signal for Automatic Caption Generation," IEICE Tech. repo. SP99-27, Vol. 99, No. 121, pp. 7-14, 1999, 6.

[3] N. Inokura, Y. Rikiso, "A Method of Synchronizing Voice Data with the Corresponding Japanese Text Automatically (in Japanese)," the IEICE trans. on info. and sys., Vol.J89-D, No.2, pp.261–270 (2006)

[4] C-W Huang, W Hsu and S-F Chang, "Automatic Closed Caption Alignment Based on Speech Recognition Transcripts," Technical Report, Columbia ADVENT,2003, 12.

[5] G Boulianne *et. al*, "Computer-assisted closed-captioning of live TV broadcasts in French," *In Proc. of INTERSPEECH-2006*, pp.273–276, 2006, 9.

[6] Y. Nishizawa, M. Sugiyama, "Correspondence between Sound and Text using Speech Features and Language Information (in Japanese)," *Proc. of the 2004 Autumn meeting of the ASJ*", 1-P-3, pp.167-168 (2004)

[7] The National Institute for Japanese Language, "Corpus of Spontaneous Japanese," http://www.kokken.go.jp/en/research_projects/kotonoha/csj/

[8] A. Lee, T. Kawahara, and K. Shikano, "Julius – an Open Source Real-Time Large Vocaburary Recognition Engine," *In Proc. of EUROSPEECH-2001*, pp.1691–1694, 2001, 9.

[9] http://julius.sourceforge.jp/en/

[10] Fukuoka city Mayer Office, "the public relation TV program of Fukuoka city (in Japanese)," http://www.pr-fukuoka-city.stream.jfit.co.jp/index.php