WORD SEQUENCE PAIR SPOTTING FOR SYNCHRONIZATION OF SPEECH AND TEXT IN PRODUCTION OF CLOSED-CAPTION TV PROGRAMS FOR THE HEARING IMPAIRED

Ichiro Maruyama[1] *Yoshiharu Abe*[1,2] *Takahiro Wakao*[1]
Eiji Sawamura[1] *Terumasa Ehara*[1,3] *Katsuhiko Shirai*[1,4]

[1]Telecommunications Advancement Organization of Japan, 3-23-5 Uehara, Shibuya, Tokyo, 151-0064 Japan
[2]Mitsubishi Electric Corporation, 5-1-1, Ofuna, Kamakura, Kanagawa, 247-8501 Japan
[3]NHK, 1-10-11 Kinuta, Setagaya, Tokyo, 157-8510 Japan
[4]Waseda University, 3-4-1 Okubo, Shinjuku, Tokyo, 169-8555 Japan

ABSTRACT

This paper describes a method of automatically synchronizing TV news speech and its captions. A news item consists of sentences and often has a corresponding computerized text, which can be used as a caption. We have developed a new phonetically HMM-based word spotter. In this word spotter, word sequences before and after a synchronization point are concatenated and scoring is based on the state of the synchronization point.

The detection accuracy of the proposed method is shown to be superior to a conventional method using no word sequence pair. Model configurations are shown for detection failure, an announcer's misstatements and restatements, and erroneous transcriptions. A 100% detection rate with no false alarms is achieved by combining multiple word sequence pairs in series. A 100% detection rate with few false alarms is obtained by using model configurations for misstatements or erroneous transcriptions.

1. INTRODUCTION

Television is indispensable to human life in the modern world. However the people who are seeing or hearing impaired cannot enjoy TV programs as much as they want to. In closed-caption service, the speech sound in TV programs is transcribed and superimposed on TV pictures for the benefit of the hearing impaired. Although this kind of service is available for more than 70% of the TV programs in the United States, it is only available for 10% of the TV programs in Japan. Currently in Japan the closed-captions are manually produced and it is a time-consuming and costly task. Thus Telecommunications Advancement Organization (TAO) of Japan, with the support of the Ministry of Posts and Telecommunications, has initiated a project, in which electronically available texts of TV news programs are automatically synchronized with the speech and video, then superimposed on the original programs for the closed-caption service [1,2].

In synchronization work, since the announcer does not always correctly read the news text the HMM Viterbi algorithm is considered to be unsuitable for finding the correspondence between the speech and text [4]. On the other hand, word spotting would be applicable to this kind of synchronization since the correct text is already known. Thus we have developed a new phonetically HMM-based word spotter, called a word sequence pair model and have tried to achieve a detection rate of 100% with few false alarms.

Jeanreanaud et al. [3] reported that the performacne of a phonetically HMM-based word spotter depends greatly on its configuration. Therefore we will present and discuss model configurations using multiple word sequence pairs in a news text.

There are problems that can occur in the practical use of a word sequence pair model since an announcer sometimes mistates or restates the text. In addition, a news text may not be converted correctly into a transciption. We therefore propose and disscuss the model configurations to solve these problems.

2. SYNCHRONIZATION OF SPEECH AND TEXT USING WORD SEQUENCE PAIR SPOTTING

First, the automatic transcribing system with morphological analysis converts a TV news text written in Japanese into a stream of phonetic transcriptions. Second, a word sequence pair model that is based on a phonetically HMM-based word spotter [3] is generated. As shown in Figure 1, a word sequence pair model is comprised of two word sequences, which exist before and after the synchronization point in the news text. The score at state B is calculated, and we search for a local maximum in this score profile in order to determine the timing for superimposing the text.

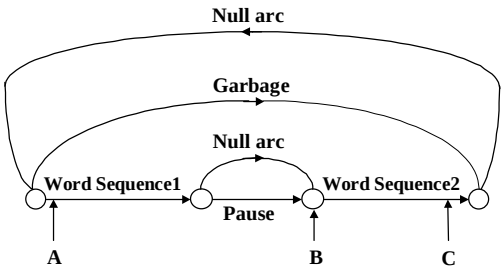

Figure 1: Word sequence model (basic model).

We use the forward-backward algorithm to estimate the probability. The score $\gamma_t(w)$ for word sequence pair w at time t is given using the forward probability $\alpha_t(i)$ for state i at time t and the backward probability $\beta_t(i)$ for state i at time t, as:

$$\alpha_t(i) = \sum_h \alpha_{t-1}(h)\, a_{hi}\, b_i(o_t) \quad (1)$$

$$\beta_t(i) = \sum_j a_{ij}\, \beta_{t+1}(j)\, b_j(o_{t+1}) \quad (2)$$

$$\gamma_t(w) = \Pr(s_t = e_B \mid O, \lambda) = \frac{\alpha_t(e_B)\,\beta_t(e_B)}{\sum_{all_s} \alpha_t(s)\,\beta_t(s)} \quad (3)$$

Where e_B is the head state of word sequence2, and the summation in equation (3) is taken over all the states s.

A garbage model consists of all Japanese phonetic HMMs. Each word sequence is expressed by concatenating the HMMs. A null arc and a pause model are inserted for the breath timing that corresponds to the individual announcer's breathing patterns for easy understanding.

3. WORD SEQUENCE PAIR MODEL CONFIGURATIONS

In this section we investigate the model configurations by combining multiple word sequence pairs in a news text. We also present model configurations for an announcer's misstatements and restatements and for the erroneous transcriptions by the automatic transcribing system.

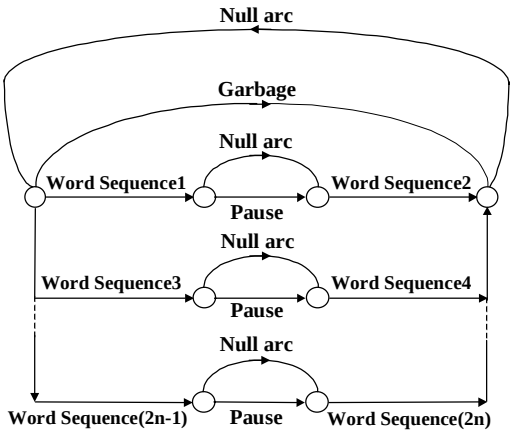

Figure 2: Parallel model.

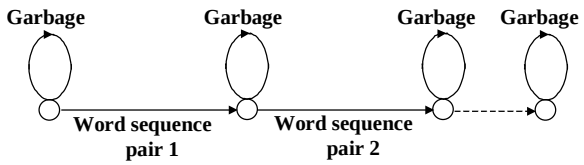

Figure 3: Series model.

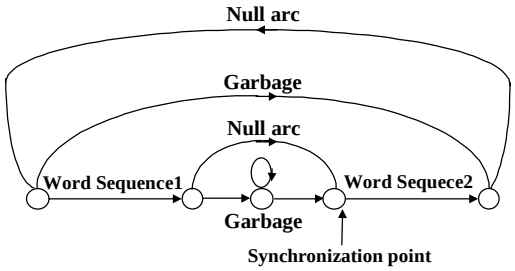

Figure 4: Misstatement and restatement model.

3.1. Model Using Multiple Word Sequence Pairs

Parallel Model

The parallel model configuration as shown in Figure 2 is comprised of a garbage model and multiple word sequence pairs in a news text. This model is expected to absorb the non-synchronizing speech with both the garbage model and the other word sequence pairs, which are not in the synchronization point.

Series Model

In the series model configuration as shown in Figure 3, word sequence pairs in a news text are concatenated in series and each garbage model is inserted between them. They are inserted in written order in the news text. This model uses the context information in the news text and is expected to improve the detection accuracy.

3.2. Models for Misstatement and Erroneous Transcription

Misstatement and Restatement Model

We obtained parts of misstatements and restatements from the actual TV news program database (DB1), which contains ten hours of recording which is taken by TAO in 1997. We found that about 90% of the misstatements and restatements occurred between phrases (*bunsetsues*). We propose a misstatement and restatement model as shown in Figure 4. This model can absorb a misstatement part with the garbage model between word sequences, and can prevent the score at the synchronization point from decreasing.

Unknown Word Model

A TV news text, which is written in *kana* and *kanji*, is converted into a stream of phonetic transcriptions by the automatic transcribing system. In some cases, the morphological analyzer treats a specific *kanji* word as an undefined word because its dictionary does not include the word. As a solution of this problem, we propose an unknown word model that substitutes an unknown word for a garbage model.

Multiple Transcriptions Model

One of the errors of the automatic transcribing system, is the case where it changes a TV news text into an incorrect stream

of phonetic transcriptions. We propose a multiple transcriptions model. It uses both the best phonetic transcription and multiple phonetic transcriptions. Each word sequence is comprised of a parallel of multiple phonetic transcriptions.

4. EXPERIMENT

In this section, we compare the detection accuracy of the word sequence pair model (basic model) to methods that calculate the score of the first or last state of the word sequence pair. Then we present the experimental results on a synchronization task using the models proposed in Section 3.

4.1. Evaluation Items

(Experiment I) Basic Model

We show the advantage of the word sequence pair model by comparing the detection accuracy when the scoring is based on states A, B, or C in Figure 1. Each word sequence is comprised of 1 phrase, so that the word sequence pair model is comprised of 2 phrases.

(Experiment II) Models Using Multiple Word Sequences Pairs

We compare the detection accuracy of the basic model, the parallel model, and the series model. The parallel model is comprised of from 4 to 6 word sequence pairs, and the serial model is comprised of 3 or 4 word sequence pairs. Each word sequence is comprised of 2 phrases, so that the word sequence pair model is comprised of 4 phrases.

(Experiment III) Models for Misstatement and Erroneous Transcription

In this experiment, we compare the proposed models with the basic model. The proposed models are generated in the following way. Each word sequence is comprised of 2 phrases, so that the word sequence pair model is comprised of 4 phrases.

In the misstatement and restatement model, a garbage model is inserted only between word sequences. In the unknown model, each part of an unknown word is manually generated, so that the unknown part occupies 10 - 15% of the phonemes of the whole word sequences. In the multiple transcriptions model, 1 of 4 word sequences is replaced by multiple phonetic transcriptions and the transition probability for each transcription is set to the same value. We comprise multiple transcriptions of three kinds of transcriptions, which consist of incorrect transcription by the automatic transcribing system, correct one, and another incorrect one.

4.2. Experimental Condition

The Japanese phoneme HMMs were trained using B and C sets from the TAO news speech database (DB2)[3], which were read and recorded in a studio by announcers. The training data were read by 4 males and 9 females, and totaled about 4.76 hours. The number of phoneme HMMs was 39, and the HMMs were gender-independent with 4-states-3-loops, left-to-right and 8 Gaussian mixtures. The acoustic analysis condition is shown in Table 1.

Sampling frequency	16 kHz
High-pass filter	$1 - 0.97z^{-1}$
Window type	Hamming window
Frame length	30 ms
Frame shift	10 ms
Feature parameter	LPC cepstrum(16th) + delta LPC cepstrum(16th) + delta logarithmic power

Table 1: Acoustic analysis condition.

In experiment I, 136 synchronization points, which consisted of 34 points for 2 males and 2 females, were selected from set A in DB2. In experiment II, 164 synchronization points including all the points in experiment I were selected from A set in DB2. 30 synchronization points for each model in experiment III were selected from relatively clean speech with no background noise in DB1. They were selected in both male and female speech.

The performance of synchronization of speech and text, i.e. the detection accuracy, was measured by the detection rate and false alarm rate. The detection rate was defined as the ratio of the number of the correctly detected points whose timing errors are within 10 frames to the number of the selected synchronization points. The false alarm rate was defined as the number of falsely detected points per word sequence pair per hour.

4.3. Experimental Results

(Experiment I) Basic Model

The results of the comparison are presented in Table 2 and Table 3. The word sequence pair model, in which scoring is based on state B, achieved the best performance in terms of detection rate, false alarm rate, and timing errors. D/R in Table 3 and in the following tables stands for detection rate.

(Experiment II) Models Using Multiple Word Sequence Pairs

The results of the comparison with the parallel model, the series model, and the basic model are shown in Table 4. In the series model, the log scores of almost all the synchronization points were 0, and it obtained the ideal result of a 100% detection rate with no false alarms. The scores of the detected synchronization points in the parallel model were higher than those in the basic model, because multiple word sequence pairs in the parallel model also work as another garbage model.

(Experiment III) Models for Misstatement and Erroneous Transcription

The results of the comparison between the misstatement and restatement model and the basic model are shown in Table 5. While in the basic model one-third of the synchronization points were not detected because the synchronization points' scores were low, the misstatement and restatement model achieved a detection rate of 100% with a false alarm rate of 2.26.

The results of the comparison of the unknown models with 1 or 2 unknown words and the basic model are shown in Table 6. A detection rate of 100% with a false alarm rate of below 2 was achieved even when using the unknown models. We found that some false alarms tend to occur when the number of phonemes in the word sequence pair is not so many.

The results for the multiple transcriptions model, the basic model with corrected transcriptions and the basic model with incorrect transcriptions, are compared in Table 7. The average log scores of the detected synchronization points in the multiple transcriptions model were higher by 140 than those in the basic model (incorrect). On the other hand, compared with the basic model (correct), almost same performance is achieved, although the average of the scores was degraded by 1.

5. CONCLUSION

In this paper we proposed word sequence pair spotting for the synchronization of speech and text and demonstrated that this method can detect the timing more accurately than methods which calculate the score of the first or last state of the word sequence pair. We concentrated on the model configurations and conducted the synchronization experiments using both the news database recorded in a studio and the actual TV news database with no background noise. The series model, in which multiple word sequence pairs are combined in series, achieved a detection rate of 100% with no false alarms. In the model configurations that can deal with an announcer's misstatements and restatements, as well as with errors created by the automatic transcribing system, a 100% detection rate with few false alarms was obtained. Further work will include the synchronization in actual TV news speech with background noise.

6. REFERENCES

1. T. Ehara, T. Wakao, E. Sawamura, I. Maruyama, Y. Abe, and K. Shirai, “Application of natural language processing and speech processing technology to production of closed-caption TV programs for the hearing impaired”, Proc. NLPRS'97, pp. 273-278, 1997

2. T. Wakao, T. Ehara, E. Sawamura, Y. Abe, and K. Shirai, “Application of NLP technology to production of closed-caption TV programs in Japanese for the hearing impaired”, ACL 97 workshop, Natural Language Processing for Communication Aids, pp. 55-58, 1997

3. P. Jeanrenaud, K. Ng, M. Siu, J. R. Rohlicek, and H. Gish, “Phonetic-Based Word Spotter: Various Configurations and Application to Event Spotting”, Proc. ESCA EuroSpeech93, pp. 1057-1060, 1993

4. A. Ando, “Automatic Synchronization of news speech with its caption”, Proc. IWHIT'97, pp. 65-70, 1997

State	A	B	C
Timing errors (ms)	25.4	16.7	22.4

Table 2: Timing Errors (when detection rate is 95 %).

Threshold (log likelihood)		-50	-100	-150	-200
State A	D/R (%)	78.6	92.6	96.3	97.7
	fa/kw/hour	4.49	13.62	61.47	227.9
State B (proposed)	D/R (%)	79.4	93.3	98.5	100
	fa/kw/hour	3.19	11.02	43.65	186.7
State C	D/R (%)	78.6	91.9	96.32	98.53
	fa/kw/hour	7.83	20.47	66.04	213.5

Table 3: Detection accuracy in states A, B, and C.

Threshold (log likelihood)		-10	-50	-100	-200	-300
Basic model	D/R (%)	34.7	65.2	82.9	95.7	99.3
	fa/kw/hour	0.00	0.34	0.99	6.55	12.68
Parallel model	D/R (%)	59.7	76.2	89.6	95.7	100
	fa/kw/hour	0.06	0.70	2.25	6.26	14.57
Series model	D/R (%)	100	------	------	------	------
	fa/kw/hour	0.00	------	------	------	------

Table 4: Detection accuracy for parallel model, series model, and basic model.

Threshold (log likelihood)		-50	-100	-150	-200	-250	-500	-750	-1000
Basic model	D/R (%)	0.0	0.0	3.3	6.6	6.6	50.0	63.3	36.6
	fa/kw/hour	0.06	0.06	0.06	0.22	0.44	6.85	42.30	70.13
Misstatement and restatement model	D/R(%)	36.6	63.3	80.0	93.3	100	------	------	------
	fa/kw/hour	0.17	0.22	0.39	0.83	2.26	------	------	------

Table 5: Detection accuracy for misstatement and restatement model and basic model.

Threshold (log likelihood)		-50	-100	-150	-200
Basic model	D/R (%)	63.3	93.3	96.6	100
	fa/kw/hour	0.00	0.00	0.06	0.06
Unknown model (occurrence: 1)	D/R(%)	73.3	86.6	100	-----
	fa/kw/hour	0.00	0.18	0.18	-----
Unknown model (occurrence: 2)	D/R(%)	80.0	93.3	100	------
	fa/kw/hour	0.12	0.23	1.75	------

Table 6: Detection accuracy for unknown word model and basic model

Threshold (log likelihood)		-100	-200	-300	-400
Basic model (incorrect)	D/R (%)	3.3	53.3	83.3	93.3
	fa/kw/hour	0.00	0.00	0.19	0.97
Basic model (correct)	D/R (%)	63.3	100	------	------
	fa/kw/hour	0.06	0.06	------	------
Multiple transcription model	D/R (%)	63.3	100	------	------
	fa/kw/hour	0.06	0.06	------	------

Table 7: Detection accuracy for multiple transcriptions model and basic model.