Proceedings of 20th International Congress on Acoustics, ICA 2010

23–27 August 2010, Sydney, Australia

Fast Calculation of Translation Model Score for Simultaneous Automatic Speech Recognition of Multilingual Audio Contents

Eri Ohmura and Hiroaki Nanjo

Graduate School of Science and Technology, Ryukoku University, Seta, Otsu 520-2194, Japan
ohmura@nlp.i.ryukoku.ac.jp

PACS: 43.72.Ne

ABSTRACT

This paper addresses automatic speech recognition (ASR) for multilingual audio contents, such as international conference recordings and broadcast news. For handling such contents efficiently, a simultaneous ASR is promising. Conventionally, ASR has been performed independently, namely, language by language, although multilingual speech, which consists of utterances in several languages representing identical meaning, is available. We previously proposed a bilingual speech recognition framework based on statistical ASR and machine translation in which bilingual ASR is performed simultaneously and complementarily. In this simultaneous recognition framework, ASR systems use not only acoustic and language model scores but also a translation model (TM) score. In this study, we investigate an efficient calculation method of TM scores. A TM score represents how a sentence corresponds to another sentence of different languages. In general, between different languages a word can be translated into various words. Moreover, word orders are different. Considering these characteristics, TM scores should be modeled statistically. In a statistical translation model, each word in source language is modeled to have a possibility to be translated into every word in target language. For instance, for the matching (alignment) of n-word sentences and m-word sentences, there are n to the m-th power word-alignments. For a strict calculation of statistical TM scores, first, we calculate the probability of each alignment and then calculate their sum. However, this calculation costs too much and is inadequate for a real-time system. In this study, we reduce the computational cost. Specifically, since for almost all alignments, their probabilities are much smaller compared with the highest alignment probability, we regard the highest alignment probability as a TM score. We compared TM score calculation methods for time and accuracy in a Japanese ASR using English information based on a bilingual recognition framework. We significantly reduced processing time for TM score calculation without any degradation of ASR accuracy.

INTRODUCTION

Based on the progress of information technologies and globalization, such large-scale multimedia contents as broadcast news and recordings of international conferences or meetings can be distributed quickly all over the world through wideband networks. For those living or working in countries without adequate second-language skills, simultaneous interpretation is required for quick understanding of such multimedia contents, and actually manual interpretation has been performed for many such contents. To make such multimedia contents more universal, captioning is significant. Especially for large-scale multimedia contents, automatic captioning is required. Automatic speech recognition (ASR) is promising for automatic captioning, and some captioning systems based on ASR have already been realized.

Conventionally, ASR-based automatic captioning systems mainly target monolingual speech. We focus on multilingual speech, which consists of utterances in several languages representing the same meaning. For the efficient use of such multilingual audio materials, an ASR strategy is required that handles them appropriately. Specifically, an ASR framework is needed in which the corresponding utterances of several languages are recognized simultaneously and complementarily.

Based on this background, we previously proposed a bilingual speech recognition framework[1] that requires ASRs of two or more languages and a calculation module of a translation model (TM) score. But the processing time is very time-consuming. In this study, we investigate fast calculation methods of TM scores.

Conventional studies of combination of ASR and machine translation (MT) have been mainly focusing on a computer assisted translation of texts[2] [3]. In such works, users who want to translate some texts do not write down but just utter, and ASR with text-based MT is performed to dictate the utterances. On the contrary, our research target is ASR with speech-based MT.

FRAMEWORK OF SIMULTANEOUS ASR OF MULTILINGUAL AUDIO CONTENTS

Framework of Statistical Speech Recognition

The orthodox statistical ASR is formulated by finding most probable word sequence W for input speech X, which is described as:

$$\hat{W} = \underset{W}{\operatorname{argmax}}\, P(W|X). \tag{1}$$

Based on Bayes's rule, $P(W|X)$ can be rewritten as:

$$\hat{W} = \underset{W}{\text{argmax}} \frac{P(W)P(X|W)}{P(X)}. \quad (2)$$

Since $P(X)$ does not affect the maximization, Eq. (1) is rewritten as:

$$\hat{W} = \underset{W}{\text{argmax}}\, P(X|W)P(W). \quad (3)$$

Speech recognition is a process for finding best word sequence $\hat{W}$ with two scores, $P(X|W)$ and $P(W)$. Here, a model that gives $P(X|W)$ is called an acoustic model, and a model that gives $P(W)$ is called a language model.

Generally, a logarithm function is adopted, and then scaling parameters α and β are introduced as follows. Here, N_w represents the number of the words of word sequence W:

$$\hat{W} = \underset{W}{\text{argmax}} \left(\log P(X|W) + \alpha \log P(W) + \beta N\right). \quad (4)$$

Framework of Statistical Machine Translation

The orthodox statistical MT is formulated by finding the most probable word sequence of target language J for a word sequence of source language E. The process is described in Eq. (5):

$$\hat{J} = \underset{J}{\text{argmax}}\, P(J|E). \quad (5)$$

Here, $P(J|E)$ is a translation score from source text E to target text J, namely, the correspondence scores of E and J. A model that gives score $P(J|E)$ is a TM.

Framework of Simultaneous ASR of Multilingual Audio Contents

Simultaneous multilingual ASR is formulated by finding the most probable word sequence of the target language for the input speeches of the target language and other languages. For example, in an ASR of Japanese utterance X using corresponding English utterance Y, word sequence $\hat{J}$ is looked for that gives maximum $P(J|X,Y)$. Fig. 1 shows an overview. The procedure is shown in Eq. (6). Here, $P(X,Y)$ is eliminated since it does not affect maximization:

$$\begin{aligned}\hat{J} &= \underset{J}{\text{argmax}} P(J|X,Y) \\ &= \underset{J}{\text{argmax}} \frac{P(J,X,Y)}{P(X,Y)} \\ &= \underset{J}{\text{argmax}} P(J,X,Y). \quad (6)\end{aligned}$$

Introducing E_m, which represents one possible English word sequence for English speech Y, Eq. (6) is described as:

$$\hat{J} = \underset{J}{\text{argmax}} \sum_m P(J,E_m,X,Y). \quad (7)$$

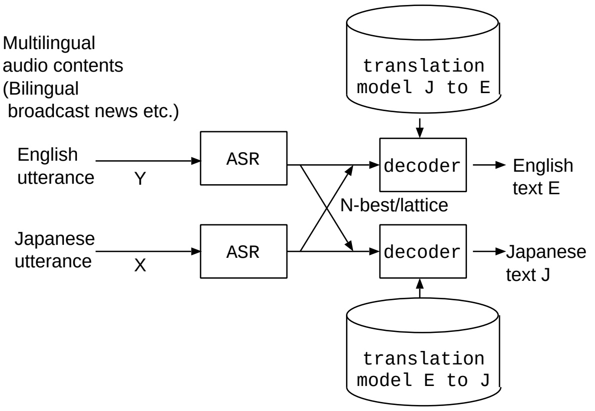

Figure 1: Overview of bilingual speech recognition framework

Here, since X and Y only depend on J and E, respectively, $P(X|J,E_m,Y)$ and $P(Y|E_m,J)$ are rewritten as $P(X|J)$ and $P(Y|E_m)$, respectively:

$$\begin{aligned}\hat{J} &= \underset{J}{\text{argmax}} \sum_m P(X|J,E_m,Y)P(J,E_m,Y) \\ &= \underset{J}{\text{argmax}} \sum_m P(X|J)P(Y|E_m,J)P(E_m|J)P(J) \\ &= \underset{J}{\text{argmax}}\, P(X|J)P(J) \sum_m P(Y|E_m)P(E_m|J). \quad (8)\end{aligned}$$

Adopting the logarithm function on the right of Eq. (8) and introducing scaling factors, Eq. (8) is rewritten as:

$$\begin{aligned}\hat{J} &= \underset{J}{\text{argmax}} \Big(\log P(X|J) + a \log P(J) + \beta N \\ &\qquad + b \log \sum_m P(Y|E_m)P(E_m|J) \Big) \\ &= \underset{J}{\text{argmax}} \Big(\log P(X|J) + a \log P(J) + \beta N \\ &\qquad + b \log \sum_m \frac{P(Y|E_m)P(E_m)P(J|E_m)}{P(J)} \Big) \\ &= \underset{J}{\text{argmax}} \Big(\log P(X|J) + (a-b) \log P(J) + \beta N \\ &\qquad + b \log \sum_m P(Y|E_m)P(E_m)P(J|E_m) \Big). \quad (9)\end{aligned}$$

Substituting $a-b$ with α and b with γ, the ASR is described as Eq. (10):

$$\begin{aligned}\hat{J} &= \underset{J}{\text{argmax}} \Big(\log P(X|J) + \alpha \log P(J) + \beta N \\ &\qquad + \gamma \log \sum_m P(Y|E_m)P(E_m)P(J|E_m) \Big). \quad (10)\end{aligned}$$

The right side of Eq. (10) consists of Japanese ASR scores in log-domain "$\log P(X|J) + \alpha \log P(J) + \beta N$", English ASR score "$P(Y|E_m)P(E_m)$", and TM score "$P(J|E_m)$".

When we use only one hypothesis for English speech Y, i.e., 1-best result, and represent the English hypothesis as E, Eq. (11) is rewritten as:

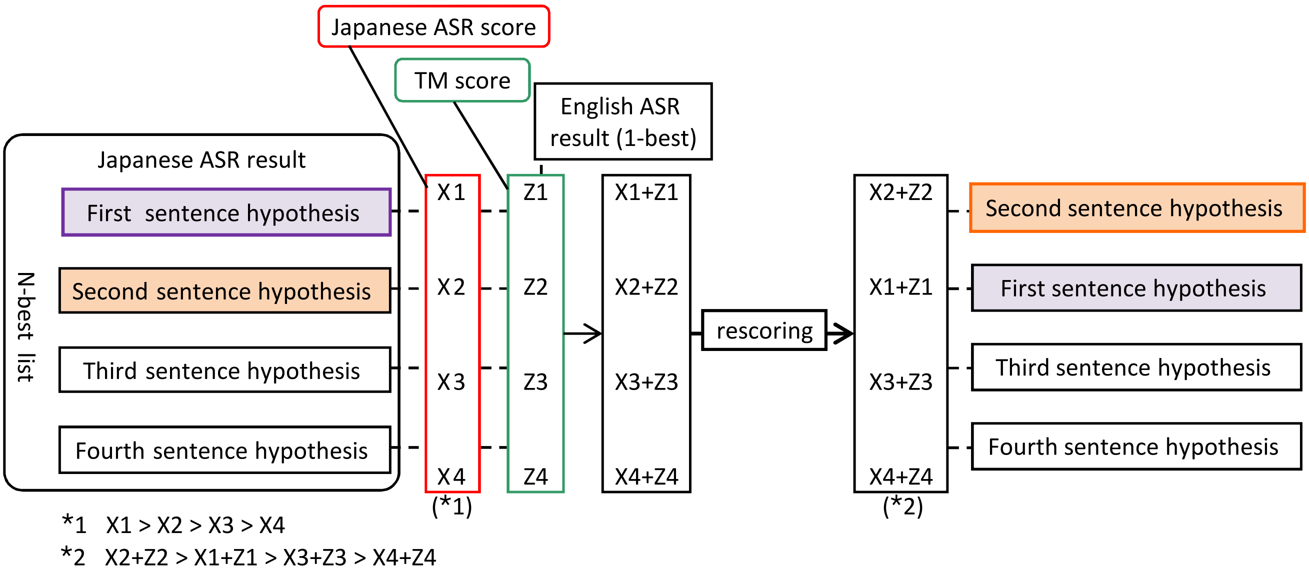
Japanese ASR score
TM score
English ASR result (1-best)
Japanese ASR result
N-best list
First sentence hypothesis
Second sentence hypothesis
Third sentence hypothesis
Fourth sentence hypothesis
X1
X2
X3
X4
Z1
Z2
Z3
Z4
X1+Z1
X2+Z2
X3+Z3
X4+Z4
rescoring
X2+Z2
X1+Z1
X3+Z3
X4+Z4
Second sentence hypothesis
First sentence hypothesis
Third sentence hypothesis
Fourth sentence hypothesis
(*1)
(*2)
*1 X1 > X2 > X3 > X4
*2 X2+Z2 > X1+Z1 > X3+Z3 > X4+Z4

Figure 2: Bilingual ASR with N-best rescoring

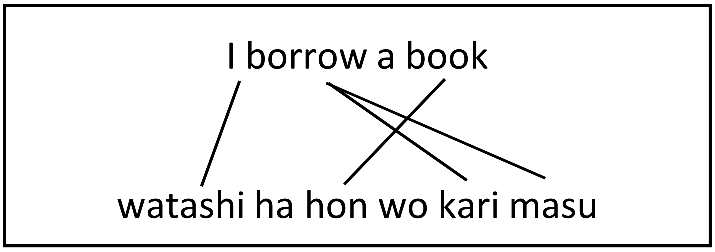

Figure 3: Example of alignment A =(1 0 4 0 2 2)

$$\hat{J} = \operatorname*{argmax}_J \Big(\log P(X|J) + \alpha \log P(J) + \beta N + \gamma \log P(J|E)\Big). \quad (11)$$

TM score $P(J|E)$ gets smaller based on the sentence length. Thus, scaling parameter δ is also introduced, and Eq. (11) is rewritten:

$$\hat{J} = \operatorname*{argmax}_J \Big(\log P(X|J) + \alpha \log P(J) + \beta N + \gamma(\log P(J|E) + \delta N)\Big). \quad (12)$$

Equation (12) shows that to perform bilingual ASR, we need three components: 1) a TM that gives the translation score, 2) an ASR system that generates ASR results and their scores, and 3) a decoder that searches for the best word sequence that maximizes a product (sum in log-scale) of the translation and ASR scores. In this study, we focus on the TM. Fig. 2 shows a concrete example.

TRANSLATION MODEL SCORE

A TM score represents how a sentence corresponds to another sentence of different languages. In general, between different languages a word can be translated into various words. Moreover, word orders are often different. Considering these characteristics, TM scores should be modeled statistically.

IBM model

In this study, we adopt the IBM model as a TM because it is one of the most widely used TMs. In a statistical TM, each word in source language is modeled to have a possibility to be translated into every word in target language.

In IBM modeling, word correspondence, that is, alignment, is introduced to calculate the correspondence score of the source and target texts. Alignment A is represented as a vector whose i-th element A_i shows that the i-th target word corresponds to the A_i-th source word. Fig. 3 shows an example of alignment A =(1 0 4 0 2 2). If A_i is 0, the target word does not correspond to any source word.

For matching n-word and m-word sentences, there are n^m matching pairs. For a strict calculation of statistical TM scores, we calculate the probabilities for all alignments and add them. For example, English text E to Japanese text J translation score $P(J|E)$ is described with alignment A (Eq. (13)):

$$P(J|E) = \sum_A P(J,A|E). \quad (13)$$

In this study, IBM models-1, -2 and -3 are used. They have the following features:

- IBM model-1: Models translation probabilities only
- IBM model-2: In addition to IBM model-1, models word positions in the source and target languages
- IBM model-3: In addition to IBM model-2, models the fertilities of words (1 to n correspondence)

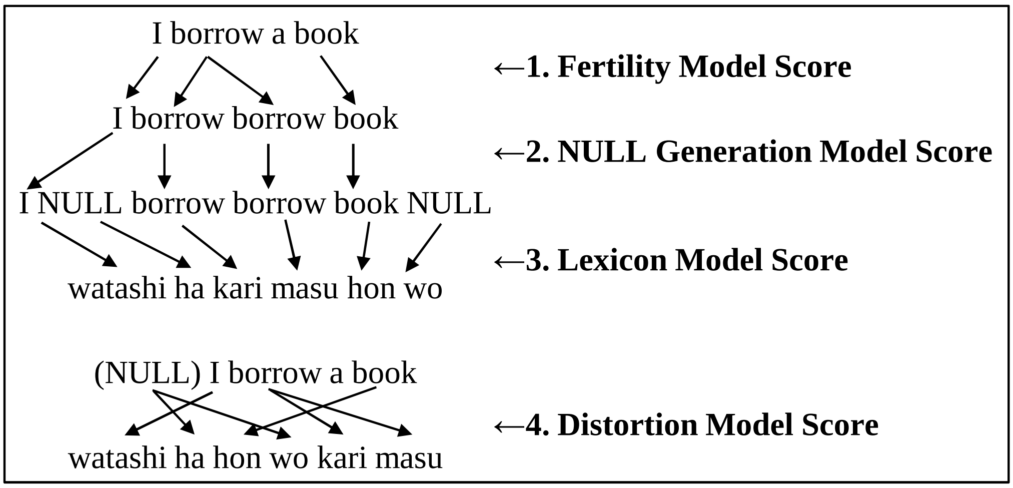
I borrow a book
←1. Fertility Model Score
I borrow borrow book
←2. NULL Generation Model Score
I NULL borrow borrow book NULL
←3. Lexicon Model Score
watashi ha kari masu hon wo
(NULL) I borrow a book
←4. Distortion Model Score
watashi ha hon wo kari masu

Figure 4: Correspondence score calculation, where alignment A=(1 0 4 0 2 2) is given

Here, IBM model-3 is explained in detail. It consists of the following four models for the calculation of $P(J|E)$ in Eq. (13):

- Fertility model: models the probability that a word in the source language corresponds to n words in the target language.
- NULL generation model: models the probability that a word in the target language does not correspond to any word in the source language.
- Lexicom model: models the probability that a word in the source language can be translated into a word in the target language.
- Distortion model: models the probability that a source word in position i moves to specific position j in target language, considering the length of the source and target sentences.

Correspondence score $P(J,A|E)$ is computable in the product of these four model scores (Fig. 4).

Next, we explain each model in detail.

Fertility model

The fertility model gives probability $n(\phi_i|E_i)$ that word E_i of a certain language corresponds to ϕ_i words in other languages. Fertility model score F_s can be shown by Eq. (14):

$$F_s = \prod_i n(\phi_i|E_i). \tag{14}$$

NULL generation model

In IBM model-3, when a target word does not correspond to any source words, dummy word "NULL" corresponds to it.

NULL generation model gives probability P_{NULL} that NULL is inserted after a certain source word. NULL generation model score N_s is shown as Eq. (15). Here, m is the number of the words of the source languages. ϕ_0 is the number of inserted NULLs:

$$N_s = P_{NULL}{}^{\phi_0} \cdot (1 - P_{NULL})^{m-\phi_0}. \tag{15}$$

Lexicon model

The lexicon model gives translation probability $t(J_i|E_{A_i})$ that word E_{A_i} of a source language is translated into word J_i of a target language. Lexicon model score L_s is shown by Eq. (16):

$$L_s = \prod_i t(J_i|E_{A_i}). \tag{16}$$

Distortion model

The distortion model gives probability $d\left(\frac{i}{j,u,v}\right)$ that the j-th word of the source sentence of u words moves the i-th word of the target sentence of v words. Here, "NULL" is assumed as the 0-th word in the source language. When alignment A is given, distortion model score D_s is Eq. (17), since $j = A_i$ by its definition:

$$D_s = \prod_i d\left(\frac{i}{A_i,u,v}\right). \tag{17}$$

Based on these four model scores, correspondence score $P(J,A|E)$ is shown by Eq. (18):

$$P(J,A|E) = F_s \cdot N_s \cdot L_s \cdot D_s. \tag{18}$$

Next, we discuss the actual calculation method of the alignment scores of J and E. The TM score is the total of $P(J,A|E)$ over all possible alignments. If source and target sentences consist of u and v words, respectively, the number of alignments is $(u+1)^v$. In this paper, we describe an efficient TM score calculation.

Specifically, $v \times (u+1)$ matrix M is generated to reduce computational cost. An example of such a matrix M is shown in Fig. 5. Here, the leftmost row is assumed to be the 0-th row. The value of element M_{ij} holds the product of the lexicon and distortion model scores, that is, $t(J_i|E_{A_i}) \times d\left(\frac{i}{A_i,u,v}\right)$. Based on the definition of the alignment, each alignment A can be generated by selecting one element for each line. If one of the selected elements is 0, TM score $P(J,A|E)$ will be 0 because the value is given by their product. Therefore for generating an alignment, we just select an element that has a value greater than 0 for each line. Actually, for many elements of matrix M, the values are 0, and many other elements have very small values that can be regarded as 0. As a result, this significantly reduces the number of alignments to be calculated from possible $(u+1)^v$ alignments.

	(NULL)	I	borrow	a	book
watashi	0	$7.25*10^{-3}$	0	0	$1.25*10^{-4}$
ha	$2.47*10^{-1}$	$1.24*10^{-2}$	0	0	$1.07*10^{-3}$
hon	0	0	$1.3*10^{-5}$	0	$1.19*10^{-2}$
wo	$3.07*10^{-6}$	0	0	$2.02*10^{-4}$	0
kari	0	0	$1.40*10^{-3}$	0	0
masu	0	0	$1.07*10^{-3}$	0	$3.14*10^{-2}$

Figure 5: Example of matrix M

Fast Calculation by Approximation of Translation Models Score

Above we described the efficient calculation method for a strict TM score calculation. The calculation of the TM model score still costs too much, so it is inadequate for a real-time system. Here, we reduce the calculation cost by the following two approximations.

First, relative threshold $THRES_{rel}$ is introduced. Specifically, the score whose value is smaller than the maximum score by $THRES_{rel}$ is regarded as 0 (Eq. (19)):

$$M_{ij} = 0 \quad \text{if} \quad \max_{j} \log M_{ij} - \log M_{ij} > THRES_{rel} \quad . \tag{19}$$

For example, in the second line in Fig. 5 indicated by "ha", the maximum value is M_{20}. When $THRES_{rel}$ is set to -2, M_{24} is regarded as 0.

Second, absolute threshold $THRES_{abs}$ is introduced. Specifically, the score that is smaller than score ($THRES_{abs}$) is regarded as 0 (Eq. (20)). In this study, $THRES_{abs}$ is set to -4:

$$M_{ij} = 0 \qquad \text{if} \qquad \log M_{ij} < THRES_{abs} \qquad . \tag{20}$$

For example, in Fig. 5, when $THRES_{abs}$ is set to -4, M_{32} and M_{40} are regarded as 0.

The simplest approximation of $\sum P(J,A|E)$ is replacing sum with maximization. Here, $P(J|E)$ is shown by Eq. (21), which is realized by selecting the maximum value for each line of matrix M:

$$P(J|E) = \max_{A} P(J,A|E). \tag{21}$$

EXPERIMENTAL RESULT

To examine the effect of the TM score calculation method, a bilingual ASR was performed. Here, to clarify the effect of the TM score calculation, we removed the influences of the ASR errors of other languages. Here, a Japanese ASR with corresponding English texts (a correct transcript, not the ASR results of English utterances) was evaluated. An ASR with this assumption corresponds to the case where there is only one hypothesis E_m in Eq. (11).

ASR System

For an acoustic model, we used a gender independent monophone model (129 stats, 64 mixtures) trained with 260 hours of speech read by 4130 speakers. They are based on a continuous density Gaussian-mixture HMM. Speech analysis was performed every 10 msec, and a 25-dimensional parameter was computed (12 MFCC + 12 Δ MFCC + Δ Power). For the language model, a word trigram model with a vocabulary of 60 K words trained with 350 M words from newspaper articles was used. We set up an ASR system that consists of these models and a decoder Julius rev.3.4.2 [4].

Table 1: Training data of translation model

	English	Japanese
# of sentences	56K	56K
# of words	1.3M	1.6M

Table 2: Evaluation data of bilingual ASR

# of utterances	250 by 5 Japanese speakers (news readings in Japanese)
# of sentences	50 (J-E aligned sentences)
# of words	711 (Japanese), 476 (English)

Translation Model and Training Data

As a translation model, we adopted the IBM model-3. We trained the translation model with bilingual texts from Reuters newspaper articles [5] in which Japanese and English sentences were aligned. In training, words that occur less than twice were regarded as unknown words. The statistics of unknown words were then used in the calculation of their translation scores. The specifications of the training data are shown in Table 1.

Evaluation data

In this paper, the bilingual ASR framework is evaluated on a read speech recognition task. The evaluation data are designed as follows. First, we selected 50 aligned Japanese-English sentences from an English textbook for Japanese learners that consists of transcriptions of broadcast news (English) and their translated texts (Japanese). Then we asked five Japanese speakers to read the Japanese parts (translated texts) and set them as test data (250 utterances). The specifications are shown in Table 2.

Result

Table 3 lists the Word Error Rate (WERs) of the Japanese ASR with and without English information. Here, for calculation based on Eq. (13), $THRES_{rel}$ was set to -2 and $THRES_{abs}$ was set to -4. We achieved a lower WER with English information than without it. These results show that a bilingual ASR is effective.

Next, we investigated an approximation of the TM score calculation in detail. We set $THRES_{abs}$ to -4 and tested the calculation times and the WERs with several $THRES_{rel}$. The results are shown in Fig. 6, where the horizontal axis represents $THRES_{rel}$, the left vertical axis represents calculation time

Table 3: WERs of Japanese ASR with and without English information

	WER
without English texts	12.58%
with English texts	12.21%

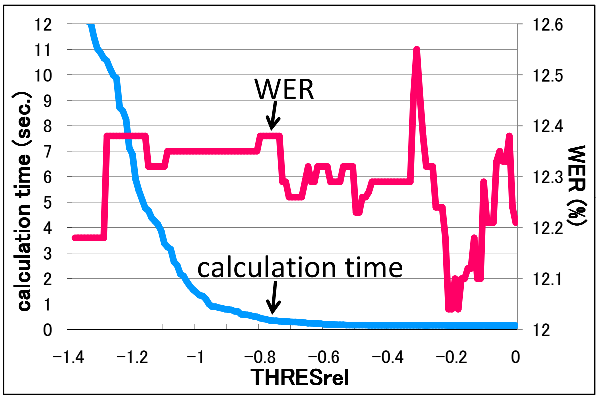

Figure 6: Calculation times and WERs

Table 4: Number of sentences for each source sentence length

# of words in a sentence	# of sentences
1-10	75
11-15	89
16-20	51
21-30	35

(seconds) of the TM score, and the right vertical axis represents the WERs of the Japanese ASR with English texts. We can reduce the calculation times based on $THRES_{rel}$. The time is steady where we used $THRES_{rel}$ higher than -0.6. Moreover, regardless of $THRES_{rel}$, that is, the calculation time, WER was stable at about 12.2 to 12.3%. Therefore, the proposed TM score approximation method effectively reduced computational cost without any degradation of recognition accuracy.

Next, the effect of the number of source words on the TM score calculations and WERs was investigated. The results are shown in Figs. 7 and 8. In Fig. 7, the horizontal axis represents $THRES_{rel}$, and the vertical axis represents the calculation time of the TM scores (seconds). In Fig. 8, the horizontal axis represents $THRES_{rel}$, and the vertical axis represents the WERs of the Japanese ASR with English texts. The figures on the graphs represent the number of words included in the source language sentence (Tabel 4). When we set $THRES_{rel}$ to -0.6 or more, we confirmed that there are no significant changes of calculation time regardless of the number of source words. When a $THRES_{rel}$ smaller than -0.6 is used, we confirmed that longer sentences require much more calculation time. We failed to confirm that even for longer sentences, strict calculation does not work well; that is, the TM calculation method is not influenced by the number of words. If we choose $THRES_{rel}$ from -0.6 to 0, we can achieve fast TM score calculation time and adequate improvement of ASR.

CONCLUSION

We investigated the fast calculation of TM scores for a simultaneous ASR of multilingual audio contents and significantly reduced the TM score calculation cost without any degradation of ASR performance. We confirmed that our proposed calculation method is effective.

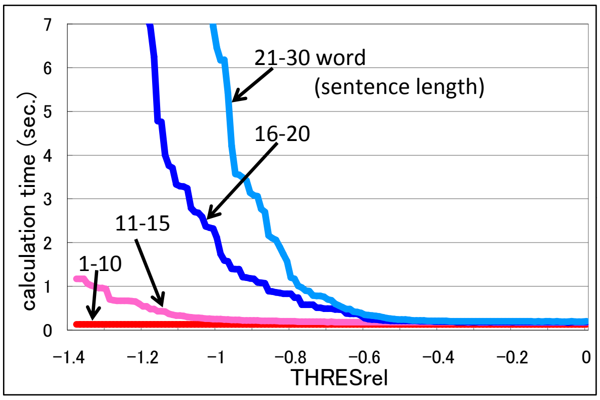

Figure 7: Calculation time for sentence length

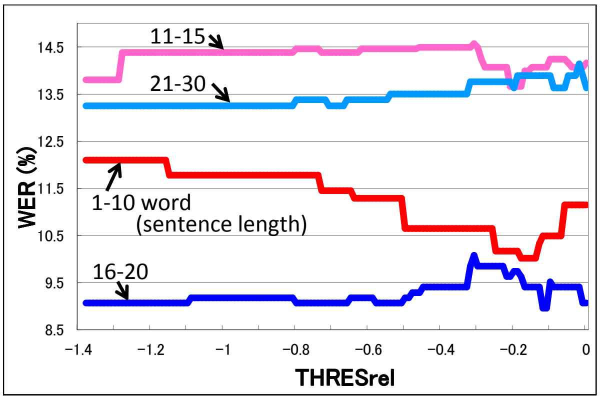

Figure 8: WERs for sentence length

ACKNOWLEDGMENT

This work was partially supported by KAKENHI (21700210).

REFERENCES

[1] H.Nanjo, Y.Oku, and T.Yoshimi, “Automatic Speech Recognition Framework for Multilingual Audio Contents,” in *Proc.INTERSPEECH*, 2007, pp. 1445–1448.

[2] S.Khadivi, A.Zolnay, and H.Ney, “Automatic Text Dictation in Computer-Assisted Translation,” in *Proc.INTERSPEECH*, 2005, pp. 2265–2268.

[3] M.Paulik, C.Fugen, S.Stuker, T.Schultz, T.Schaaf, and A.Waibel, “Document Driven Machine Translation Enhanced ASR,” in *Proc.INTERSPEECH*, 2005, pp. 2261–2264.

[4] A.Lee, T.Kawahara, and K.Shikano, “Julius – an Open Source Real-Time Large Vocabulary Recognition Engine,” in *EUROSPEECH*, 2001, pp. 1691–1694.

[5] M.Utiyama and H.Isahara, “Reliable measures for aligning Japanese-English news articles and sentences,” in *Annual meeting of Computational Linguistics 2003*, 2003, pp. 72–79.