A NEW ASR EVALUATION MEASURE AND MINIMUM BAYES-RISK DECODING FOR OPEN-DOMAIN SPEECH UNDERSTANDING

Hiroaki Nanjo† *and Tatsuya Kawahara*‡

†Faculty of Science and Technology, Ryukoku University
Seta, Otsu 520-2194, Japan
nanjo@rins.ryukoku.ac.jp

‡Academic Center for Computing and Media Studies, Kyoto University
Sakyo-ku, Kyoto 606-8501, Japan
kawahara@i.kyoto-u.ac.jp

ABSTRACT

A new evaluation measure of speech recognition and a decoding strategy for keyword-based open-domain speech understanding are presented. Conventionally, WER (word error rate) has been widely used as an evaluation measure of speech recognition, which treats all words in a uniform manner. In this paper, we define a weighted keyword error rate (WKER) which gives a weight on errors from a viewpoint of information retrieval. We first demonstrate that this measure is more appropriate for predicting the performance of key sentence indexing of oral presentations. Then, we formulate a decoding method to minimize WKER based on Minimum Bayes-Risk (MBR) framework, and show that the decoding method works reasonably for improving WKER and key sentence indexing.

1. INTRODUCTION

The major target of large vocabulary continuous speech recognition has shifted to spontaneous speech [1] [2]. For "understanding" of open-domain speech such as oral presentations and lectures, detection of important segments, namely, key sentence indexing is a promising approach. Since the orthodox key sentence indexing methods focus on keywords that are characteristic to the speeches, such keywords should be detected with higher priority by the automatic speech recognition (ASR) system.

Conventionally, speech recognition aims at perfect transcription of the utterance, and the recognition accuracy is evaluated by word error rate (WER), which is the minimum string edit distance (Levenshtein distance) between the correct transcription and the recognition hypothesis. In this framework, keywords and functional words, even fillers, are treated in a same manner. For key sentence indexing, however, keywords are apparently significant than other words. Therefore, WER is not an appropriate evaluation measure of recognition accuracy when we want to use ASR systems for speech understanding.

In previous studies on speech understanding, keyword recognition accuracy was adopted only for definite tasks such as flight information, where a set of keywords can be determined by the back-end system [3]. But it is not straightforward to define keywords for open-domain speech. In this paper, we introduce a new evaluation measure of speech recognition, that is, weighted keyword error rate (WKER) based on tf-idf criterion used in information retrieval. Then, speech recognition is designed to minimize WKER based on the Minimum Bayes-Risk (MBR) framework [4]. We demonstrate that the decoding method works reasonably for speech understanding based on key sentence indexing.

2. AUTOMATIC INDEXING OF KEY SENTENCES FOR SPEECH ARCHIVES

We address automatic indexing of key sentences, which will be useful indices of speech archives in oral presentations. Collection of these sentences may suffice summarization of the talk [5]. The framework extracts a set of natural sentences, which can be aligned with audio segments for alternative summary output.

2.1. Automatic Transcription System

First, automatic transcription system is described. For model training, we use the *Corpus of Spontaneous Japanese (CSJ)* [6] [1] which was complied by the "Spontaneous Speech Corpus and Processing Technology" project. It consists of a variety of academic presentation speeches at technical conferences and simulated public speakings on given topics. They are manually given orthographic and phonetic transcriptions.

For language model training, we use 2592 presentations whose text size in total is 6.7M words (=Japanese morphemes). A trigram language model is trained for the vocabulary of 24K words. As for acoustic model training, we use 781 presentations that amount to 106 hour speech. We constructed a gender-independent PTM (phonetic tied-mixture) triphone model [7]. Here, 129 codebooks of 192 mixture components were used. We also revised our recognition engine Julius so that very long speech can be handled without prior segmentation [8].

With adaptation of the acoustic and language models, the word error rate of 22.0% was obtained for the test-set of 15 academic presentation speeches [9].

2.2. Keyword-based Key Sentence Indexing

An orthodox key sentences indexing approach is to focus on keywords that are characteristic to the oral presentation. The most popular statistical measure to define and extract such keywords is the following tf-idf criterion.

Table 1. Human performance of key sentence indexing (50% indexing)

Presentation ID	F-measure	κ-value
A01M0007	0.756	0.566
A01M0035	0.725	0.521
A01M0056	0.605	0.321
A01M0074	0.657	0.400
A01M0097	0.622	0.355
A01M0110	0.773	0.583
A01M0137	0.742	0.552
A01M0141	0.653	0.390
A03M0016	0.585	0.303
A03M0106	0.635	0.384
A03M0112	0.821	0.669
A03M0156	0.569	0.291
A04M0051	0.748	0.551
A04M0121	0.584	0.303
A04M0123	0.688	0.467
A05M0011	0.750	0.555
A05M0031	0.758	0.566
Average	0.697	0.480

$$S_{KW}(w_j) = tf_j * \log(N_d/df_j) \quad (1)$$

Here, term frequency tf_j is the occurrence count of a word w_j in the oral presentation, and document frequency df_j is the number of oral presentations (=documents) in which the word w_j appears. N_d is the number of presentations used for normalization. For each sentence s_i, we compute $S_{KW}(s_i) = \sum_{w_j \in s_i} S_{KW}(w_j)$.

Then, key sentences are selected based on the score up to a specified number (or ratio) of sentences from the whole presentation.

2.3. Evaluation Measure of Key Sentence Indexing

For evaluation of key sentence indexing, we use 17 academic presentations of the CSJ which are listed in Table 1. For these presentations, texts are manually segmented into sentence units based on a fixed guideline, and then key sentences are labeled by three human subjects. The subjects were instructed to select sentences which seemed important by 50% of all.

We prepared answer sets based on the agreed portion of the 50% extraction data for reliable and meaningful evaluation. Specifically, we picked up sets of sentences agreed upon by two subjects. Since three combinations exist for picking up two subjects out of three, we derived three answer sets. The performance is evaluated by averaging for these three sets. Using this scheme, we can also estimate the human performance by matching one subject's selection with the answer set derived from the other two. The recall, precision, F-measure and κ-value are 81.9%, 60.6%, 0.697 and 0.480, respectively as shown in Table 1 and Table 2. Here, F-measure is a normalized mean of recall and precision rates and κ-value is often used to measure agreement by considering the chance rate. These figures are regarded as a target for the automatic indexing system.

2.4. Result of Key Sentence Indexing

We conducted an evaluation of indexing using the transcriptions generated by the ASR system. Table 2 lists the recall, precision rates and F-measure in comparison with the case of manual transcription. Since the derived sets of sentences for automatic and manual transcription are different, we automatically align the hypothesized sentences with the correct ones, and calculate accuracy based on the alignment.

Table 2. Results of key sentence indexing from ASR results

	transcript.	indexing	recall	precision	F-measure
(1)	manual	manual	81.9%	60.6%	0.697
(2)	manual	auto	70.8%	52.5%	0.603
(3)	auto	auto	71.1%	44.2%	0.545

Comparing the indexing performance by the system against human judgment with manual transcription (cases (1) and (2) in Table 2), the accuracy is lower by about 15%. The indexing method works reasonably, but it still has room for improvement. Comparing the cases (2) and (3) in Table 2, it is observed that the ASR degraded accuracy, especially on the precision. In [10], however, we showed that major cause is incorrect sentence segmentation by automatic period insertion rather than word substitution errors.

3. EVALUATION MEASURES OF SPEECH RECOGNITION FOR KEY SENTENCE INDEXING

3.1. Generalization of Word Error Rate

Word error rate (WER) is widely used to evaluate ASR accuracy. It is defined as equation (2). Here, *N* is the number of words in the correct transcription, *S* is the number of substitution errors, *D* is the number of deletion errors, and *I* is the number of incorrectly inserted words (insertion errors).

$$\text{WER} = \frac{I + D + S}{N} * 100 \quad (2)$$

For each utterance, DP (Dynamic Programming) matching of the recognition result and the correct transcription is performed to identify the correct words and calculate WER.

Apparently, in WER, all words are treated in a uniform manner or with a same weight. However, there must be a difference in the weight of errors, since several "keywords" have more impact on understanding of the speech than trivial functional words. Thus, the WER is not optimal evaluation measure when we want to use ASR systems for speech understanding, which includes key sentence indexing.

Based on the background, we generalize WER and introduce Weighted Word Error Rate (WWER), in which each word has a different weight according to its influence on the speech understanding. WWER is defined as follows.

$$\begin{aligned}
\text{WWER} &= \frac{V_I + V_D + V_S}{V_N} * 100 && (3)\\
V_N &= \Sigma_{w_i} \quad v_{w_i} && (4)\\
V_I &= \Sigma_{\hat{w}_i \in I} \quad v_{\hat{w}_i} && (5)\\
V_D &= \Sigma_{w_i \in D} \quad v_{w_i} && (6)\\
V_S &= \Sigma_{seg_j \in S} \quad v_{seg_j} && (7)\\
& v_{seg_j} = \max(\Sigma_{\hat{w}_i \in seg_j} v_{\hat{w}_i}, \Sigma_{w_i \in seg_j} v_{w_i}) && (8)
\end{aligned}$$

Here, v_{w_i} is a weight of word w_i, which is the i-th word of the correct transcription, and $v_{\hat{w}_i}$ is a weight of word $\hat{w}_i$, which is the i-th word of the ASR result. And seg_j represents the j-th

ASR result	:	a	b	c	d	e	f	
Correct transcript	:	a		c	d'		f	g
DP result	:	*C*	*I*	*C*	*S*		*C*	*D*

$$\text{WWER} = (V_I + V_D + V_S)/V_N * 100$$

$$V_N = v_a + v_c + v_{d'} + v_f + v_g$$
$$V_I = v_b$$
$$V_D = v_g$$
$$V_S = \max(v_d + v_e, v_{d'})$$

v_i: weight of word *i*.

Fig. 1. Example of weighted word error rate (WWER) calculation

substituted segment and v_{seg_j} is a weight of segment seg_j. For the segment seg_j, total weight of the correct words and total weight of the recognized words are calculated, and then v_{seg_j} is set to a larger one. In this work, we use alignment for WER to identify the correct words and calculate WWER. Thus, WWER is equivalent to WER if all word weights are set to 1. In Fig. 1, an example of WWER calculation is shown.

3.2. Weighted Keyword Error Rate

In this work, we use the ASR system for key sentence indexing, which is the first step of speech understandings. As described in section 2.2, we adopt the tf-idf measure for key sentence indexing. Therefore, as a weight of word w_j, we use its tf-idf value $S_{KW}(w_j)$. Here, tf_j is calculated using the N-best list, and at this point, it is different from $S_{KW}(w_j)$ in equation (1).

In the indexing process, keywords are selected from nouns which do not include proper nouns, pronouns and numbers, and only keywords have tf-idf value (= word weight). In this case, we assume that non-keywords have zero weight. WWER calculated with these assumptions is then defined as weighted keyword error rate (WKER). Keyword error rate (KER), which is calculated by setting all keyword weights to 1, is also used for comparison.

3.3. Relation between ASR Evaluation Measures and Key Sentence Indexing Accuracy

Then, we analyzed the correlations of the ASR evaluation measures with the performance of key sentence indexing. The same 17 oral presentations shown in Table 1 are used for the analysis. For each presentation, 10 cases of speech recognition were conducted with several language models, acoustic models and decoding parameters (insertion penalty)[1], and 170 recognition results were generated for correlation analysis. Here, we use normalized F-measure $N(F)$ and κ-value $N(\kappa)$ as evaluation measures of key sentence indexing. They are defined as the system performance normalized by the human performance so that the human performance is 1 for every presentation.

Results are listed in Table 3. It is confirmed that the proposed measure WKER has the highest correlation with indexing accuracy. On the other hand, WER and KER were not significantly correlated with indexing accuracy, especially for normalized F-measure. These facts show that WKER is more appropriate for predicting the performance of key sentence indexing.

[1]Task matched/unmatched models and speaker independent/dependent models are used.

Table 3. Relation between ASR evaluation measures and indexing evaluation measures

	$N(F)$	$N(\kappa)$
Word Error Rate (WER)	0.00	0.28**
Weighted WER (WWER)	0.09	0.29**
Keyword Error Rate (KER)	0.14	0.37**
Weighted KER (WKER)	0.20**	0.40**

$N(F)$: normalized F-measure
$N(\kappa)$: normalized κ-value
**: significantly correlated (1%)

4. MINIMUM BAYES-RISK DECODING FOR SPEECH UNDERSTANDING

Since we confirmed the correlation between WKER and key sentence indexing accuracy, in this section, we explore a decoding strategy to minimize WKER. It is based on the Minimum Bayes-Risk (MBR) framework [4].

4.1. Concept

The orthodox statistical speech recognition is formulated as finding the most probable word sequence $\hat{W}$ for an input speech X, which is described in equation (9).

$$\hat{W} = \underset{W'}{\text{argmax}}\ P(W'|X) \tag{9}$$

In the Bayesian decision theory, ASR is described with a decision rule $\delta(X) : X \rightarrow \hat{W}$. Using a real-valued loss function $l(W, \delta(X)) = l(W, W')$, the decision rule minimizing Bayes-Risk is given as follows [4].

$$\delta(X) = \underset{W}{\text{argmin}} \sum_{W'} l(W, W') \cdot P(W'|X) \tag{10}$$

It is equivalent to the orthodox speech recognition described in equation (9) when the 0/1 loss function is used in equation (10). In our baseline ASR system, this decoding is used.

In order to minimize WER, Levenshtein distance, which is equivalent to WER, is conventionally used as a loss function $l(W, W')$ [4] [11]. In this work, we want to minimize the weighted keyword error rate (WKER) to improve key sentence indexing accuracy, thus we define the loss function based on WKER as described in equation (11).

$$\delta(X) = \underset{W}{\text{argmin}} \sum_{W'} \text{WKER}(W, W') \cdot P(W'|X) \tag{11}$$

Since $P(W'|X)$ can be rewritten as $P(W', X)/P(X)$ and $P(X)$ does not affect the minimization, equation (11) is rewritten as follows.

$$\delta(X) = \underset{W}{\text{argmin}} \sum_{W'} \text{WKER}(W, W') \cdot P(W', X) \tag{12}$$

Table 4. Result of WKER minimization decoding

ID	WKER 1-best→ MBR	Key sentence indexing accuracy (F-measure) 1-best→ MBR
A04M0123	42.19 → 42.20	0.555 → 0.535
A04M0121	41.19 → 41.16	0.482 → 0.498
A04M0051	9.38 → 8.85	0.630 → 0.630
A01M0056	10.66 → 9.73	0.529 → 0.532
A01M0035	41.90 → 40.67	0.533 → 0.526
A01M0007	8.91 → 8.76	0.600 → 0.600
A01M0110	35.15 → 35.19	0.629 → 0.629
A01M0141	26.31 → 26.39	0.508 → 0.484
A01M0137	43.56 → 42.93	0.537 → 0.545
A01M0074	34.47 → 33.89	0.508 → 0.522
A01M0097	4.09 → 3.15	0.547 → 0.533
A03M0112	15.90 → 14.95	0.506 → 0.507
A03M0016	38.03 → 36.99	0.564 → 0.551
A03M0156	39.76 → 38.11	0.446 → 0.452
A03M0106	53.64 → 52.60	0.485 → 0.479
A05M0011	49.83 → 49.39	0.558 → 0.585
A05M0031	22.57 → 22.48	0.621 → 0.639
Average	25.57 → 24.96	0.545 → 0.548

Table 5. Comparison of decoding methods

minimization target	WER	WKER	key sentence indexing accuracy (F-measure)
WER	25.69	25.00	0.545
WKER	26.10	24.96	0.548
baseline	25.94	25.57	0.545

Moreover, a normalizing parameter λ is also adopted [4], so the decision rule is finally described as follows.

$$\delta(X) = \underset{W}{\text{argmin}} \sum_{W'} \text{WKER}(W, W') \cdot P(W', X)^{\frac{1}{\lambda}} \qquad (13)$$

To find the best word sequence W in a practical way, an N-best list is generated by the baseline ASR system, and then N-best rescoring is performed.

4.2. Result

We evaluated WKER minimization decoding and its effect for key sentence indexing using the same test-set (17 presentations). For each utterance, we generate N-best list with $N = 1000$. The rescoring parameter λ is set to 18 based on preliminary experiments.

Table 4 shows the result of WKER minimization decoding. The proposed decoding strategy improved WKER from 25.57% to 24.96%. The result verifies that it worked properly as designed. Table 5 lists the average improvement of WER, WKER and key sentence indexing accuracy achieved by WKER minimization decoding in comparison with the conventional MBR decoding. According to the WKER improvement, the key sentence indexing accuracy is also improved to 0.548. On the contrary, when MBR decoding is performed to minimize WER instead of WKER, WER reduction was achieved, but there is no improvement for indexing accuracy. It is confirmed that WKER minimization decoding works reasonably for improving keyword-based indexing of oral presentations.

5. CONCLUSION

We first addressed the ASR evaluation measure in terms of speech understanding of open-domain, and introduced WKER based on the criterion for information retrieval. Then, we designed a decoding strategy to minimize WKER. It is shown that WKER is an appropriate measure and WKER minimization decoding is effective from the viewpoint of speech understanding.

Acknowledgment: The work was conducted in the Science and Technology Agency Priority Program on "Spontaneous Speech: Corpus and Processing Technology". The authors are grateful to Prof. Sadaoki Furui and other members for the collaboration in this fruitful project.

6. REFERENCES

[1] S.Furui, K.Maekawa, and H.Isahara, "Toward the realization of spontaneous speech recognition – introduction of a Japanese priority program and preliminary results –," in *Proc. ICSLP*, 2000, vol. 3, pp. 518–521.

[2] S.Furui, "Recent advances in spontaneous speech recognition and understanding," in *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003, pp. 1–6.

[3] T.Kawahara, C.-H.Lee, and B.-H.Juang, "Flexible speech understanding based on combined key-phrase detection and verification," *IEEE Trans. Speech & Audio Process.*, vol. 6, no. 6, pp. 558–568, 1998.

[4] V.Goel, W.Byrne, and S.Khudanpur, "LVCSR rescoring with modified loss functions: A decision theoretic perspective," in *Proc. IEEE-ICASSP*, 1998, vol. 1, pp. 425–428.

[5] I.Mani and M.Maybury, Eds., *Advances in Automatic Text Summarization*, MIT Press, Cambridge, 1999.

[6] K.Maekawa, "Corpus of Spontaneous Japanese: Its design and evaluation," in *Proc. ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003, pp. 7–12.

[7] A.Lee, T.Kawahara, K.Takeda, and K.Shikano, "A new phonetic tied-mixture model for efficient decoding," in *Proc. IEEE-ICASSP*, 2000, pp. 1269–1272.

[8] T.Kawahara, H.Nanjo, and S.Furui, "Automatic transcription of spontaneous lecture speech," in *Proc. IEEE Workshop on Automatic Speech Recognition and Understanding*, 2001.

[9] H.Nanjo and T.Kawahara, "Language model and speaking rate adaptation for spontaneous presentation speech recognition," *IEEE Trans. Speech & Audio Process.*, vol. 12, no. 4, pp. 391–400, 2004.

[10] H.Nanjo, T.Kitade, and T.Kawahara, "Automatic indexing of key sentences for lecture archives using statistics of presumed discourse markers," in *Proc. IEEE-ICASSP*, 2004, vol. 1, pp. 449–452.

[11] A.Stolcke, Y.Konig, and M.Weintraub, "Explicit word error minimization in N-best list rescoreing," in *Proc. EUROSPEECH*, 1997, pp. 163–165.