A Metric for Evaluating Speech Recognizer Output based on Human-Perception Model

Nobuyasu Itoh[1], *Gakuto Kurata*[1], *Ryuki Tachibana*[1], *Masafumi Nishimura*[2 †]

[1] IBM Research, Toyosu, 135-8511, JAPAN,
[2] Graduate School of Informatics, Shizuoka University

(iton,gakuto,ryuki)@jp.ibm.com, nisimura@inf.shizuoka.ac.jp

10.21437/Interspeech.2015-321

Abstract

Word error rate or character error rate are usually used as the metrics for evaluating the accuracy of speech recognition. These are naturally-defined objective metrics and are helpful for comparing recognition methods fairly. However the overall performance of the recognition systems and the usefulness of the results are not necessarily considered. To address this problem, we study and propose a metric which replicates human-annotated scores using their perception to the recognition results. The features that we use are the numbers of insertion errors, deletion errors, and substitution errors in the characters and the syllables. In addition we studied the numbers of consecutive errors, the misrecognized keywords, and the locations of errors. We created models using linear regression and random forest, predicted human-perceived scores, and compared them with the actual scores using Spearman's rank-based correlation. According to our experiments the correlation of human perceived scores with character error rates is 0.456, while those with the predicted scores by using a random forest of 10 features is 0.715. The latter is close to the averaged correlation between the scores of the human subjects, 0.765, which suggests that we can predict the human-perceived scores using those features and that we can leverage human perception model for evaluating speech recognition performance. The important factors (features) for the prediction are the numbers of substitution errors and consecutive errors.

Index Terms: speech recognition, evaluation, word error rate, character error rate, human perception

1. Introduction

Evaluation metrics are important for comparing various algorithms for speech recognition and assessing which approach outperforms the others. The WER (Word Error Rate) has been most often used in research work on speech recognition [1]. The CER (Character Error Rate) is also popular in research for some languages (such as Japanese), where the word units can be ambiguous [2]. In some applications, task-oriented metrics have also been tried. For example Levit proposed a metric for end-to-end accuracy evaluation in voice-enabled search tasks [3]. In addition, speech-enabled dialog systems have been usually evaluated by the concept error rate, task completion ratio, or required time and average number of turns to the completion [4][5], because the quality of dialog systems is affected not only by the recognition accuracy but also by interpretation algorithms and strategies of the dialog management. However, in many speech applications such as dictation, voicemail transcription, or message creation, the quality of the output text, (its readability and understandability), is the largest factor for evaluating the speech recognition systems. In this paper we study the human-assessed quality of Automatic Speech Recognition (ASR) output to rank the transcriptions in Japanese where the orthographic and word units are hard to define.

This paper is organized as follows: in Section 2 we survey related works. In Section 3 the strategy for collecting human perceived scores from human annotators is described. In Section 4, we discuss the features to be used for defining our metrics. Section 5 presents our experimental results using two frequently used approaches: linear regression and random forest, a well-known non-linear prediction method. In Section 6, we discuss the results, followed by some concluding remarks.

2. Related works

Jones [6] studied the readability of ASR transcripts, and reported that certain metadata (capitalization, punctuation, and disfluencies (removed or not)) significantly influenced the human-perceived readability. Nanjo [7] proposed the Weighted Key word Error Rate (*WKER*) as a more suitable metric for evaluating ASR used in applications. The ASR quality problem is similar in difficulty to assessing (or automatically predicting) human reactions to machine translation output . BLEU was proposed as a more objective metric for evaluating machine translation [8]. However, the relation between the metrics for speech recognition output and human-perceived quality has not been sufficiently investigated. We are investigating what and how the recognition errors affect human perception of the ASR output. Our goal is to the predict human perceived quality scores using surface features in the ASR output, then to leverage the prediction model for evaluating ASR.

[†] This work was conducted when the author was a member of IBM Research - Tokyo

3. Strategy for data sampling and protocol

3.1. ASR output sample

In order to investigate the human perception of quality, the method for choosing the samples of ASR output is quite important. The human-perceived quality also depends on the task domains, because human subjects usually compensate for missing information by using their own background knowledge. To avoid the influences, we focused on a daily business mail, where the topics should not require any specific knowledge, as our target domain. Here is a sample text:

> *今日 十七時 大丈夫ですか*
> *Is seventeen o'clock today OK for you*

All of the fillers are removed and no punctuation is uttered or automatically inserted. All of the numerical expressions appear in Kanji characters.

The lengths of uterances and the distribution of the recognition accuracy are also important control parameters in these kinds of experiments. Context is also a key to how well we can understand messages. With neither of context nor reference such as a single-word output, it is impossible to evaluate the quality. Overly garbled text is also hard to rank on the basis of quality and is usually classified into the "terrible or lowest-ranked" category. In contrast too-many perfect samples are also inappropriate to our objective. We therefore selected samples of ASR output to be presented to our human annotators based on these two rules:

- Length should be at least 10-characters
- The sample ASR output was divided into three subsets on the basis of %CER (character error rate) and :

1. %CER < 20: 25% of samples
2. 20 ≤%CER < 35: 50%
3. 35 ≤%CER < 50: 25%

Our pool of ASR output was created by IBM Attila recognizer [9] using a general purpose acoustic model and an open-domain language model for Japanese. The vocabulary size is about 500K.

The sample text set presented to each participant includes 16 standard (the same for all participants) samples and 32 different samples. The average length of samples was 14.1 characters.

3.2. Experimental protocol

The experimental protocol, scoring scale, score definitions, and instructions for the participants significantly affect the results. Jones [6] asked participants to give readability scores on a scale of 1-7. The ratings also depend on whether or not an accurate reference text. We prepared instructions for participants, where the rule based criterion on the scale of 0-5 are described. Table 1 shows the descriptions.

The participants were asked to evaluate each ASR output twice; once without any reference, and again after listening to the audio corresponding to the text. The first evaluation simulates a reader or receiver of the text message. The second evaluation simulates a user (speaker) of the speech recognizer. 5 participants are university students or researchers. Before starting, they tried to evaluate several other samples and were allowed to ask questions. The evaluation time was not limited, but all of them completed the evaluation task in about 30 minutes. A total of 240 (48 ×5) evaluated samples were obtained.

Table 1 *Descriptions of the scoring*

Before listening

Scores	Look at the text of the ASR output
0	Complete message or one with just slight errors, easy to understand
1	Understandable without special efforts
2	Understandable with additional efforts, but not easy, requiring some guessing
3	Difficult to understand
4	Meaningless message, just guessing
5	Worse than 4. Even guessing is difficult which might lead to misunderstanding

After listening

Scores	Compare the text of ASR output with the audio recording
0	Exactly the same message as was heard
1	With slight differences that do not have any impact on the interpretation
2	Correct summary or key part of the message (ASR output)
3	With errors significantly impact on the meaning
4	Not helpful for interpreting the utterance. Partial guessing is possible
5	Worse than 4. Even guessing is difficult, or significant misunderstanding found

4. Features and methodology

4.1. Baseline metrics and features

We conducted some preliminary experiments and selected the features to be used for metric calculations. The baselines are two popular metrics for Japanese speech recognition; Character Error Rate (*CER*) and Syllable Error Rate (*SyER*). The errors consist of substitutions (*Sub.*), insertions (*Ins.*), and deletions (*Del.*), and the error rate is calculated as

(# of *Sub.* + # of *Ins.* + # of *Del.*) / (# of characters (or syllables) in reference text),

where # represents some number. The ratios of *Sub.*, *Ins.*, and *Del.* of characters and syllables can be used as separate features (*SubR*, *InsR*, and *DelR*, call the tuple of them *3CER* (character based) and *3SyER* (syllable based) respectively).

One of the major difficulties in evaluating Japanese ASR output is how to handle synonyms with exactly the same pronunciation, which are commonly used (i.e. アイ・ビーエム, IBM). One method for handling this is to provide a list of exchangeable synonyms. We created a reasonably useful list.

4.2. Additional features

Keyword error rate

Some important words tend to affect human perception more than other words. In Japanese the content words, such as nouns and verbs, are said to be more important. We selected them in the reference text and counted the rates of three types of character errors in those keywords. (*KeySubR*, *KeyInsR*, *KeyDelR*, call the triplet of them *Key3CER* (character based) and *Key3SyER* (Syllable based) respectively)

Consecutive errors

Consecutive errors are supposed to be more influential than separate errors of the same counts. We track the longest string of consecutive errors in characters (*MaxLenCE*) and in syllables (*MaxLenSyE*) for these features, ignoring error types.

Positional errors

The beginning word in each message is usually important for understanding the entire message. In Japanese tail words also play major roles in conveying tense and modality. We define positional error features (*Init* and *Last*) as

$$Init = \begin{cases} 0 \ (Beginning\ word\ is\ correct) \\ 1 \ (Begining\ word\ is\ misrecognized) \end{cases}$$

$$Last = \begin{cases} 0 \ (Tail\ \ word\ is\ correct) \\ 1 \ (Tail\ \ word\ is\ misrecognized) \end{cases}$$

4.3. Methodologies

Many approaches for prediction are available from rule based approaches to machine learning. We used two approaches in our predictions of the human ratings. One approach is linear regression, while the other is non-linear, the random forest approach which achieved one of the best results for categorization problems [10].

5. Experimental results

We conducted experiments comparing the baseline metrics and predicted qualities of the ASR output with human perceptions of the quality scores. The predicted quality was obtained from trained models using features, which can be regarded new model-based metrics for evaluating ASR. The comparisons used Spearman's correlation coefficient [11], which is better for comparing how closely correlated two rank-based scores are. Since the amount of data (the number of samples) was not large, we used the leaving one-out technique to compare each model-based metric with the others. First we selected one sample for testing, all of the other samples were used for training, this model was used to predict the quality score for the testing sample. The process was repeated for each selected testing sample and we ultimately obtained metric values for all of the samples for comparison. The number of trees in random forest approach was 50. The ratio of Out-of-Bag data is 1/3.

Table 2 shows the distribution of scores before and after listening. We see that many after-listening scores are better than the before-listening ones. The human annotators initially believed that something was wrong with those texts, but after listening to the audio, realized that they were more acceptable or even quite accurate in some cases, where syllable sequences greatly assisted their understanding.

Table 2 *Comparison of before and after listening*

		After listening					
		0	1	2	3	4	5
Before listening	0	53	4	4	3	0	0
	1	35	7	8	7	1	0
	2	10	11	10	11	4	1
	3	2	4	8	23	6	3
	4	0	0	0	2	10	4
	5	0	0	0	0	2	7

Table 3 shows the correlations between the metrics using various features and human-perceived scores. Overall the correlations with before-listening scores are better than those with after-listening scores. In our intuition the reference audio is helpful for better ratings. This result might seem inconsistent. Unnatural character sequences degrade the readability and make it harder to interpret. But the interpretation was correct, where some errors do not necessarily affect humans' interpretation thanks to their compensation capability. The after-listening scores were better and not correlated with objective metrics. This is a persuading story. We interviewed human annotators about their ratings, investigated the ASR texts, and noticed quite a few samples with unnatural use of Kana. These cases support John's observations that meta data affect the readability, before-listening scores. We will study more in the future.

As expected %CER is not necessarily good metric from human-perception point of view. Bleu is worse. In the perception models most of the selected features were helpful for improving the correlations. However character-based keywords error rates didn't contribute to improvements for either (before and after) listening scores, and the syllable-based keyword error rate was helpful for only after-listening (RF2, RF2a). Consequently the best metrics differ for before and after- listening scores. Metric RF5 was the best for before-listening rating. In contrast metric RF6 (with Key3SyER) had the best after-listening results. The metric 'Human' refers to the correlation among different human annotators for the samples. The performances of the best model-based metrics (RF5, RF6) are similar to 'Human', which suggests that the selected features accurately reflected the human perceptions.

We checked the weights on each feature in RF5 and RF6. Syllable *SubR*, *MaxLenCE*, and character *SubR* followed by *MaxLenSyE* were top four predictor variables for before-listening scores. Character and syllable *SubR*s, syllable *KeySubR* and *MaxLenCE* were the top four ones for after-listening scores.

6. Concluding remarks

This paper introduced novel metrics on the basis of trained model using human perceived scores, which achieved close

Table 3 *Comparison of metric*

Metric	**Features**	**Spearman's correlation**	
		Before	After
CER	N.A.	0.456	0.368
Bleu	1,…,3-grams of characters	0.414	0.307
SyER	N.A.	0.498	0.474
LinearReg1	*CER+SyER*	0.525	0.477
LinearReg2	*3CER+3SyER*	0.559	0.499
RF1	*3CER+3SyER*	0.689	0.587
RF2	*3CER+3SyER+ Key3CER+Key3SyER*	0.686	0.618
RF2a	*3CER+3SyER+Key3SyER*	0.682	0.625
RF3	*3CER+3SyER+ MaxLenCE+ MaxLenSyE*	0.714	0.620
RF4	*3CER+3SyER+Key3SyER+MaxLenCE+ MaxLenSyE*	0.697	0.634
RF5	*3CER+3SyER+ MaxLenCE+ MaxLenSyE+Init+Last*	**0.715**	0.622
RF6	*3CER+3SyER+Key3SyER+ MaxLenCE+ MaxLenSyE+Init+Last*	0.702	**0.638**
Human	N.A.	0.765	0.645

LinearReg: Linear Regression RF: Random Forrest

performances with human-to-human correlations. Unexpectedly, %CER is not well correlated with human perception, and even less with after-listening scores. It suggests that Japanese are more tolerant to spelling errors than to acoustic (syllable) errors. It is still unknown that this result is common in other languages. We will pursue better metrics in other languages. The results of model-based evaluation are encouraging, which was close to human-to-human agreements.

7. Acknowledgements

Authors would like to thank Mr. Teppei Ohno (Internship student from Toyohashi University of Technology at the time) for his hard work on our experiments and are also thankful to Dr. D. Willet, N. Ueffing (Nuance Communications), R. Bakis, and B. Ramabhadran (IBM T. J. Watson Research Center) for their helpful discussions on metrics for evaluating ASR.

8. References

[1] E. Geoffrois, "Speech Recognition Evaluation," ELRA HLT Workshop on Evaluation, (2005).

[2] M. Nishimura, N. Itoh, K. Yamasaki, "Word-based approach to Large-vocabulary Continuous Speech recognition for Japanese, Trans. of IPSJ, Vol. 40, No. 4, pp. 1395-1403, (1999).

[3] M. Levit, S. Chang, B. Buntschuh, N. Kibre, "End-to-End Speech Recognition Accuracy Metric for Voice-Search tasks," Proc. of ICASSP 2012 , (2012).

[4] J. F. Allen, B. W. Miller, E. K. Ringger, T. Sikorsky, "A Robust System for Natural Spoken Dialogue," Proc. of 34th Annual Meeting of the ACL, (1996).

[5] L. Lamel, "Spoken Language Dialog System Development and Evaluation," ISSD '98, (1998).

[6] D. Jones, F. Wolf, E. Gibson, E. Williams, E. Fedrenko, D. Reynolds, M. Zissman., "Measuring the Readability of Automatic Speech-to-Text Transcripts," Proc. of Eurospeech, pp. 1585–1588, (2003).

[7] H. Nanjo, and T. Kawahara, A New ASR Evaluation Measure and Minimum Bayes-risk Decoding, Proc. of ICASSP 2005, (2005).

[8] K. Papineni, S. Roukos, T. Ward, and W. J. Zhu, "BLEU: a method for automatic evaluation of machine translation," ACL-2002: 40th Annual meeting of the Association for Computational Linguistics, pp. 311–318, (2002).

[9] S. F. Chen, B. Kingsbury, L. Mangu, D. Povey, G. Saon, H. Solotau, and G. Zweig, "Advances in Speech Transcription at IBM Under the DARPA EARS Program, IEEE Trans., Audio Speech and Language Processing, Vol. 14, No. 5, pp. 1596-1608, (2006).

[10] R. Caruana, A. Niculescu-Mizil, "An Empirical Comparison of Supervised Learning Algorithms," 23rd ICML, (2006).

[11] http://en.wikipedia.org/wiki/Spearman's_rank_correlation_coefficient