Predicting Human Perceived Accuracy of ASR Systems

Taniya Mishra, Andrej Ljolje, Mazin Gilbert

AT&T Labs-Research
180 Park Avenue
Florham Park, NJ

taniya@research.att.com, alj@research.att.com, mazin@research.att.com

Abstract

Word error rate (WER), which is the most commonly used method of measuring automatic speech recognition (ASR) accuracy, penalizes all types of ASR errors equally. However, humans differentially weigh different types of ASR errors. They judge ASR errors that distort the meaning of the spoken message more harshly than those that do not. Aiming to align more closely with human perception of ASR accuracy, we developed a new metric HPA (Human Perceived Accuracy) that predicts the subjective perceived accuracy of ASR transcriptions. HPA is computed based on the central idea of differential weighting of different ASR errors. Applied to the particular task of automatically recognizing voicemails, we found that the correlation between HPA and the human judgement of ASR accuracy was significantly higher (r-value=0.91) than the correlation between WER and human judgement (r-value=0.65).

1. Introduction

Word Error Rate (WER) is the most commonly used method for measuring Automatic Speech Recognition (ASR) accuracy. WER is the edit distance[1] between the reference word sequence (i.e., the offline human transcription of the spoken message) and its automatic transcription by the ASR system, normalized by the length of the reference word sequence. In WER, all words are considered equally important, and all errors (substitutions, deletions, insertions) are considered equally bad.

However, in practice, the impact of all errors is not the same. If an error does not hurt a person's ability to understand a message, such an error has a much lower impact than an error that distorts the meaning of the message. Consumers are often unhappy with high impact errors (e.g., incorrect recognition of names and phone numbers) but are fairly tolerant of low impact errors (e.g., substituting words by their (perfect or even near) homophones; deletions/substitutions of function words like *an* and *the*). This is because consumers are not really measuring word-for-word accuracy between the spoken message and its transcript; rather, they are judging whether the transcript produced by ASR captures the meaning of the spoken message.

Given that ultimately humans have to comprehend the ASR transcription of the spoken message, subjective ratings (like Mean Opinion Scores; henceforth MOS) of accuracy produced by human subjects are perhaps the most accurate way to determine ASR performance. But MOS testing is expensive and time-consuming, and thus considered unrealistic to include in the day to day incremental development of any ASR engine.

Our goal is to develop a metric that can predict the human perceived accuracy of ASR transcriptions as shown by its strong correlation with mean opinion score (MOS) ratings, measures how much of the semantically significant information in the spoken message, i.e., its meaning, is captured by ASR-generated transcriptions, like humans do, and requires no specialized expert labeling other than simple transcription. With this aim, we developed the Human Perceived Accuracy (HPA) metric. HPA is so named because it is intended to predict how humans will perceive the accuracy of the text produced by ASR.

HPA measures how much of the semantically significant information contained in the spoken message is captured by the ASR transcription. Following the central hypothesis that humans differentially weigh different types of ASR errors, in HPA, we assign different weights to different types of errors in different types of words. Errors in information bearing words such as names (of people, places, etc.), phone numbers, urls, email addresses and alphanumeric sequences contribute more to the overall error than errors in function words. Within a particular type of word, different types of errors are also weighed differently. For example, substituting a word by its homophone or near homophone is penalized less than substitution by a completely different word. Or, errors in name sequences may be considered one word at a time, but number sequences are considered as a whole because an error in one digit makes the entire number sequence useless. The relative weights of different errors have been determined using a perceptual experiment involving human judgment of ASR-generated transcripts.

In this paper, HPA has been applied to the particular task of converting voicemails to text in our newest visual Voice-Mail-to-Text (VMTT) transcription system [1] that converts a conventional voice mail into formatted text following the standard punctuation, capitalization and presentation conventions and applying grayscaling to lower the impact of the words recognized with low confidence scores. The VMTT transcription system is similar to AT&T's previous system for automatically converting voicemails to text, Scanmail [2], in that the main focus of the VMTT transcription system is accurate comprehension of the key points of the spoken message, though several of the UI features differ in the two systems.

2. Related Work

The limitations of WER in measuring the accuracy of a speech recognition system and possible alternatives to it have been the subject of several past studies. Past research at AT&T Labs [3] has shown that the alignment between WER and spoken language understanding is not linear. Applied to the task of call routing, the authors interpolated the ASR word n-gram with n-grams containing phrases salient to the call-routing task

[1] Edit distance can be defined as the minimum number (or weighted sum) of insertions, deletions, amd substitutions needed to transform one string into another.

$$\boxed{HPA = 100 - \sum\{saliency\text{-}weight * (wi * insertions + wd * deletions + ws * substitutions)\}/N} \tag{1}$$

Table 1: Example demonstrating the difference between WER and HPA values.

	Transcription	**(100 -WER)**	**HPA**
Reference	Hey Steve, it's Chad. It's just before 12:00 on Monday. Give me a holler when you get a chance. And call me on my cell or at my office. Thank you.	NA	NA
Hypothesis 1	Hi Steve, it's Chad. It's just before 12:00 on Monday. Give me a holler when you get a chance. Call my cell. Thank you.	75%	78.15%
Hypothesis 2	Hey Shane, it's Brad. Just before shelf on day. Give me a holler when you transferred back. phone. Bye . Keep.	40.6%	34.5%

and found that a slight reduction in WER produced substantial changes in understanding accuracy.

The divergence between WER and understanding is shown to be even more drastic in Wang et al. 2003 [4]. Keeping understanding accuracy as the main objective, the authors proposed using a language model that was obtained from an example-based language algorithm that optimized understanding accuracy. In their task, they found that though doing so increased word error rate, the overall understanding error was reduced significantly. Garfalo et al. 1999 [5] and Grangier et al. 2003 [6] also demonstrated the lack of tight alignment between WER and understanding in the spoken document retrival task. They showed that significant changes in WER does not necessarily produce significant changes in retrieval performance.

Several researchers have proposed alternatives to the WER metric. Morris et al. 2004 [7] take an information theoretic approach and propose two metrics that compute the proportion of information communicated or lost by automatically transcribing the spoken message. A similar information-based approach is also presented in McCowan et al. 2004 [8].

A number of alternatives for measuring ASR performance are explored in Garfalo et al. 1998 [9]. As in our approach, the metrics proposed by Garfalo et al. were based on the central idea that information bearing words should be given greater weight in computing ASR error. Applied to the particular task of document retrieval, their proposed metrics included named entity WER, stop-word-filtered WER, (IR-filtered) stemmed-stop-word-filtered WER, and salient query-word error rate. Significant correlations between each of these metrics and the eventual document retrieval performance were found.

The aforementioned related work show that neither the recognition of the need for alternatives to WER nor the development of information-based WER-alternatives is novel. However, the novelty of our work lies in that HPA is one of the few, if not the first, metric that has been developed by directly regressing on subjective ratings (MOS ratings) of ASR performance.

3. Data

To develop HPA, a set of 27 voicemails were selected from our fairly large database of voicemails such that the errors in the salient words and the non-salient words in the ASR-produced transcript of each of the voicemails were systematically varied. Salient words are information bearing words such as names (of people, places, etc.), phone numbers, urls, email addresses, alphanumeric sequences and other semantically significant words such as negations (*don't, can't*), and exceptions (*except*) etc.

A possible complication in developing a metric of ASR performance that differentially weighs erroneous words of differing information content is that it requires expert word-level information-content annotations in the reference transcripts. Since this is both time-consuming and expensive — and thus somewhat impractical for the development of real ASR systems, Garfalo et al. 1998 [7] cite this as a disadvantage of the information-content based metrics that they proposed.

To avoid this complication and develop a metric that requires no expert annotations other than simple word transcription, we estimated the saliency of the words using Inverse Document Frequency (IDF), an often used measure of word saliency in information retrieval and text mining. For each word in the corpus, IDF measures its global importance (or saliency) over the entire corpus. Our particular method of measuring saliency was this: In the corpus of voicemail messages (which may be the same one used for training the language models underlying the ASR system), we considered each voicemail to be a document, and calculated the IDF value of each of the unique words in this database of voicemails. We plotted the distribution of the IDF values of the unique words. This distribution has two tails, a tail of rarely occurring words (high IDF values) and a tail of frequently occurring words (low IDF values). We retrieved the words that are in these tails by retrieving the words that have IDF values 2 standard deviations below (or above) the mean of the IDF distribution. This captured the 2.1% (approx.) most frequently occurring (or, most rare) words in the voicemail corpus. We considered the tail of frequently occurring words to be the low saliency words and the words in the tail of rarely occurring words to be the high saliency words. By using this IDF-based approach for identifying high and low saliency words, we avoided relying on human expert annotation to identify salient words, thus bypassing the expensive and time-consuming process, which tends to bottleneck the evaluation process.

A point of concern in using this IDF-based approach for identifying high and low saliency words is that words such as "not", "no", "don't" and other negations, which can completely alter the meaning of a message, may occur so frequently in the corpus that their IDF values are low enough to push them into the tail of low saliency words. To prevent this, we filtered the low-saliency list to eliminate any such negation words.

In this work, erring on the side of caution, we only distinguished between the low saliency words and everything else. That is, all the words except for the 2.1% most frequently occurring words were considered to be high saliency words.

Another concern may be that since IDF values are corpus-dependent, salience ranking of words using IDF values may be unstable. But interestingly, it has been shown in [10] that (under the assumption that term usage patterns remain stable in the corpus notwithstanding changing corpus size) IDF values computed on random samples of a fair-sized corpus have strong associations to the global IDF values. On average, IDF computed on a 10% random sample can explain 80% of the variance in the global IDFs. The same work also found that terms with high document frequencies (that is, the low saliency words) remain stable to changes in corpus size.

Since we only distinguish the 2.1% lowest saliency words from all other words in the corpus, we expect that this subset of words will remain reasonably stable to any increase in corpus size, as will any regression weights that were computed for the HPA metric based on this distinction. Staying within the domain of voicemail messages, we expect the stable term usage patterns assumption to hold in our corpus.

4. Web-based Mean Opinion Scores Study

To determine the relative weights of the different types of ASR errors considered in developing the HPA metric, we conducted a web-based perceptual experiment to obtain subjective ratings of the ASR-rendered transcripts of the 27 selected voicemails. In this test, we presented the human subjects with 27 stimuli pairs, each pair consisting of a voicemail and its corresponding ASR output. A random ordering of the 27 voicemail-ASR transcript pairs was presented to each subject. For each pair, the subjects were instructed to *first read* the ASR transcript, *and then listen* to the voicemail. Then, they were instructed to imagine that they were only given the ASR output and to answer the question: Presented alone, how useful would this transcription be? using the given five-point MOS scale, where 0 indicated “not useful” and 4 indicated “very useful”.

50 adults participated in this web-based perceptual experiment. We selected the scores of the 44 individuals who completed the entire experiment. The scores obtained from each of these individuals were z-normalized and scaled to lie on a 0 to 100 scale. Then, the 44 subjects’ scores for each voicemail-ASR transcript pair were averaged to produce an average subject score for that voicemail-ASR transcript pair.

For each voicemail-ASR transcript pair, the average subject score is a measure of the average human perceived accuracy of the transcript produced by the ASR system. We consider it to be a measure of perceived accuracy even though the subjects were not directly asked about accuracy but rather about usefulness of each transcript because a transcript is only useful to the extent that it accurately captures the meaning of the voicemail. We avoided asking subjects directly about accuracy because we wanted them to evaluate not the one-to-one mapping between the spoken voicemail and the associated ASR transcript but rather the extent to which the transcript captured the relevant information in the message; i.e., the meaning of the message.

5. Regression Analysis

Having obtained the average subject score per voicemail-ASR transcript pair, we performed multiple linear regression analysis to predict the average subject score, using the following different types of errors that the ASR system made in converting a spoken voicemail message into text: 1) Errors in low saliency words, 2) Errors in high saliency words, 3) Insertion errors, 4) Deletion errors, 5) Perfect homophone substitution errors, 6)

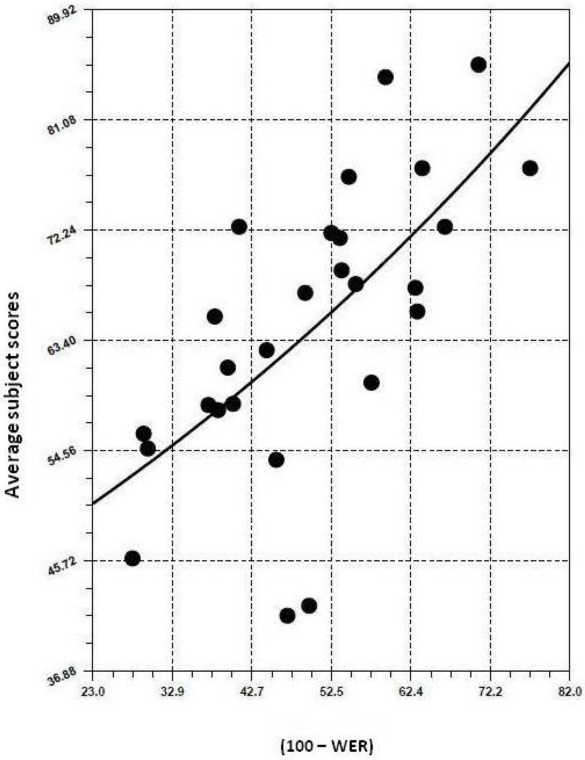

Figure 1: *Correlation of (100 - WER) to average subject scores (adjusted r value = 0.65).*

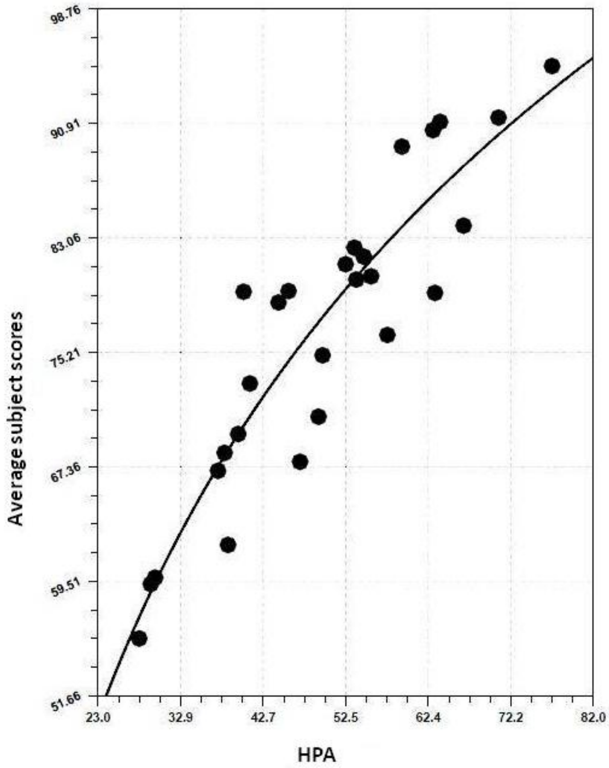

Figure 2: *Correlation of HPA to average subject scores (adjusted r value = 0.91).*

Near homophone substitution errors, 7) All other substitution errors. None of these aforementioned types of errors required any specialized human annotation. All classes of errors were automatically obtained by comparing the reference transcription of a voicemail against the ASR-generated hypothesis.

Word saliency was obtained using our IDF-based approach. To determine if a word in the human annotation of a voicemail message and a word in the same position in the ASR-generated hypothesis were perfect homophones, we did a lookup in a precompiled paired-list of perfect homophones. To determine if a pair of words in the human annotation and the ASR-generated hypothesis of the same voicemail message were "near" homophones, we compared the soundex codes of the pairs of words.

The regression was set up so that the average subject score prediction was constrained to lie on a 0 to 100 scale. On this scale, 0 indicates that people will be able to get none of the semantically relevant information from the ASR-generated transcript of the spoken voicemail message, i.e., the human perceived accuracy of the ASR-produced transcript is 0%; while 100 indicates that people will be able to get all of the semantically relevant information from the ASR-generated transcript of the spoken voicemail message, i.e., the human perceived accuracy of the ASR-generated transcript is 100%.

We set up the regression equation defining the human perceived accuracy (HPA) metric as Eq. 1. In this equation, *saliency-weight* indicates how much the error in a particular word contributes to the total error, depending on whether it is a high-saliency word or a low-saliency word; *wi* is the weight of insertion errors; *wd* is the weight of deletion errors; and *ws* is the weight of the different types of substitution errors. These weights were determined by regressing on the average subject scores obtained from our perceptual experiment.

6. Results

To estimate how well HPA can predict how humans will perceive the accuracy of the text output by ASR, we computed the correlation between HPA and the average subject scores for the ASR-generated transcripts of the 27 voicemail messages used for the experiment. We also wanted to estimate how well WER, which is the standard metric of accuracy used in speech recognition, predicts how humans will perceive the accuracy of the ASR text output. However, since WER and HPA are directionally opposed metrics (in WER, the desired value is 0, while in HPA , the desired value is 100), to compare HPA to WER, we directionally parallelized WER to HPA by computing ASR accuracy as (100 - WER). Having done so, we computed the correlation between (100 - WER) and the average subject scores for the same set of voicemails.

In computing these two correlations, we systematically experimented with thirty different linear and non-linear regressions to fit the relationship between x (the objective metric under consideration) and y (the normalized and scaled mean subject scores indicating perceived accuracy of the voicemails).

The best fit correlations obtained are as follows:

(100 - WER) and average subject scores: Best correlated by an exponential function (Fig. 2). The correlation coefficient is 0.65, i.e., it only explains about 42% of the variance in the mean subject scores.

HPA and average subject scores: Best correlated by a log function (Fig. 3). The correlation coefficient is 0.91, i.e. it explains about 83% of the variance in the mean subject scores. Both correlations are statistically significant at $p < 0.0001$.

The results show that (1) the proposed metric HPA correlates well with mean subject scores, and (2) the correlation between HPA and mean subject scores is substantially stronger than the correlation between WER and mean subject scores.

An example demonstrating how HPA and WER values differ is shown in Table 1. Notice the difference between the HPA and the (100 - WER) values for each of the two hypothesis for the same reference string. In hypothesis 1, the HPA value is higher than the (100 - WER) value. This is because most of the salient words are correctly recognized in this ASR-generated transcript. On the other hand, in hypothesis 2, the HPA value is lower than the (100 - WER) value. This is because most of the correctly recognized words in this hypothesis have lower salience or lower information content.

7. Conclusions and Future Work

We have proposed a new metric, HPA, to predict human perceived accuracy of ASR systems. Evaluations indicate that HPA may be able to predict human perceived ASR accuracy quite well, and perhaps better than the commonly used WER. To be more certain of whether HPA indeed predicts human perceived ASR accuracy better than WER, in the future we intend to use the regression equation defining HPA as a predictive model and have a separate heldout set for measuring correlations.

HPA differentially penalizes different types of ASR errors in different types of words depending on their contribution to distorting the meaning of the spoken message, but requires no more expert labeling than that needed for WER, which is simple transcription. In the future, we intend to apply HPA to other ASR applications, conduct further MOS studies, and explore other methods for capturing word saliency beyond raw IDF.

8. References

[1] Ljolje, A., Goffin, V., Caseiro, D., Mishra, T., and Gilbert, M., "Visual voice mail to text on the iPhone/iPad". Submitted to Interspeech '11, 2011.

[2] Bacchiani, M., Hirschberg, J., Rosenberg, A., Whittaker, S., Hindle, D., Isenhour, P., Jones, M., Stark, L., and Zamchick, G., "SCANMail: Audio navigation in the voicemail domain," in Proceedings HLT '01, 2001.

[3] Riccardi, G. and Gorin,A. L., "Stochastic language models for speech recognition and understanding," in Proceedings of ICSLP, Sidney, Australia, 1998.

[4] Wang, Y. Y., Acero, A., and Chelba, C., "Is word error rate a good indicator for spoken language understanding accuracy," in Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), December 2003.

[5] Garofolo, J., Auzanne, C., and Voorhees, E., "The TREC spoken document retrieval track: A success story," in Proceedings of the Eighth Text REtrieval Conference (TREC 8), November 1999.

[6] Grangier, D., Vinciarelli, A., and Bourlard, H., "Information retrieval on noisy text," IDIAP-COM 03-08, IDIAP, 2003.

[7] Morris, A., Maier, V., and Green, P., "From wer and ril to mer and wil: improved evaluation measures for connected speech recognition," in Proceedings of the International Conference on Spoken Language Processing, 2004.

[8] McCowan, I., Moore, D., Dines, J., Gatica-Perez, D., Flynn, M., Wellner, P., Bourlard, H., "On the use of information retrieval measures for speech recognition evaluation," Technical Report IDIAP-RR 04-73, Martigny, Switzerland, 2004.

[9] Garofolo, J., Voorhees, E., Auzanne, C., Stanford, V., and Lund, B., "1998 TREC-7 spoken document retrieval track overview and results," in Proceedings of TREC 7, 1998.

[10] Fu, X., and Chen, M., "Exploring the stability of IDF term weighting," AIRS 2008, pp. 10–21.