Optimizing Speech Recognition Evaluation Using Stratified Sampling

Janne Pylkkönen, Thomas Drugman, Max Bisani

Amazon

{jannepyl, drugman, bisani}@amazon.com

Abstract

Producing large enough quantities of high-quality transcriptions for accurate and reliable evaluation of an automatic speech recognition (ASR) system can be costly. It is therefore desirable to minimize the manual transcription work for producing metrics with an agreed precision. In this paper we demonstrate how to improve ASR evaluation precision using stratified sampling. We show that by altering the sampling, the deviations observed in the error metrics can be reduced by up to 30% compared to random sampling, or alternatively, the same precision can be obtained on about 30% smaller datasets. We compare different variants for conducting stratified sampling, including a novel sample allocation scheme tailored for word error rate. Experimental evidence is provided to assess the effect of different sampling schemes to evaluation precision.

Index Terms: speech recognition, evaluation, stratified sampling, bootstrapping

1. Introduction

Being able to reliably evaluate automatic speech recognition (ASR) performance is essential for developing and monitoring such systems. The data to evaluate on must be chosen such that it is as realistic and representative as possible, and it must be transcribed carefully. To minimize the manual work involved in the process, there is an urge to carry out the evaluations with as little data as possible, as long as evaluation reliability is not compromised.

In this paper, evaluation reliability or *precision* refers to the fluctuation of the extracted metric, caused by the random effects of sampling the data from the overall population. Sampling is explicit when we have a large pool of data, and only a part of it is chosen for manual labeling. Sampling can also occur implicitly, e.g. when a limited set of users participate in data collection for evaluation purposes. This paper concentrates on the former: we assume that there is an abundance of data, but labeling it is the bottleneck in the evaluation process.

Evaluation set size plays a critical role in evaluation reliability. The more data there is for evaluation, the better estimates we can obtain for the true performance of the system. When designing data collection and manual labeling, we should have an idea about how accurate the estimates should be. By analyzing the existing evaluation data, it is possible to state the variance or deviation which the evaluation metric exhibits. Such an analysis can help in choosing the appropriate size for the evaluation sets.

As the main method for improving ASR evaluation precision, we investigate using stratified sampling [1, 2], with model probability scores guiding the partitioning of the data. The same idea has been applied earlier in other machine learning tasks [3, 4], but here we show concretely how to apply the method in ASR, using real data and models to quantify the benefits of the method. We also derive a new sample allocation scheme to optimize word error rate evaluations.

We start in Section 2 by describing the ASR task and metrics used for the analysis. Section 3 describes how to quantify the reliability of the evaluation. In Section 4 we introduce stratified sampling and different variants used in this paper. Experimental results on applying stratified sampling for ASR are provided in Section 5. Conclusions are provided in Section 6.

2. Evaluation task and the model

This paper studies how to improve the precision of evaluation metrics in ASR. In particular, we will analyze the evaluation of a production-level German media voice search system. To conduct the analysis and simulations, we use an internal dataset of about 90000 transcribed utterances (about 60h of speech), from over a thousand speakers. The evaluation data had not been used for training the ASR models.

2.1. Metrics for speech recognition accuracy

In this study, we use two metrics typical for assessing ASR accuracy: Sentence Error Rate (SER) and Word Error Rate (WER). The former is simply the ratio of the number of erroneous utterances to the total number of utterances. WER provides a finer resolution for the errors by counting the number of word-level edit operations (insertions, deletions, and substitutions) needed for changing the true transcription to match the observed recognition output. Unlike SER, WER cannot be computed as an average over the WERs of the utterances in the evaluation set. Instead, one needs to sum all the errors of the dataset and reference transcription lengths separately, and compute WER as their ratio.

2.2. Model confidences and error rates

The ASR system used for this paper produces utterance-level *confidence* [5, 6] values in range 0–1. They represent the model's view of the probability that the recognition output is correct. An utterance-level confidence is computed as a product of word-level confidences, which again are derived from a confusion network. Similar to [7], a piecewise polynomial mapping is applied to word posterior probabilities to improve their usage as word-level confidences.

As we intend to use the utterance confidences to partition the data for stratified sampling, their distribution and relation to recognition errors are of interest. Figure 1 shows that in our case, the confidence values are in fact distributed very unevenly: most of the utterances are recognized with a high confidence. On the other hand, Figure 2 illustrates that our confidence values correlate very well with the true error rates. Figure 2 also shows the standard error of the utterance-level SERs in 10 uniformly distributed confidence bins. The shape of the standard

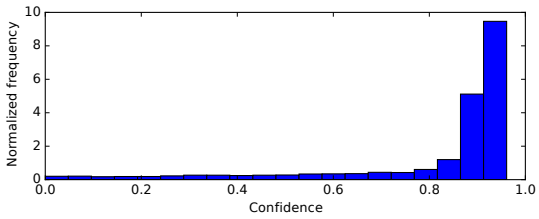

Figure 1: Histogram of utterance confidence values.

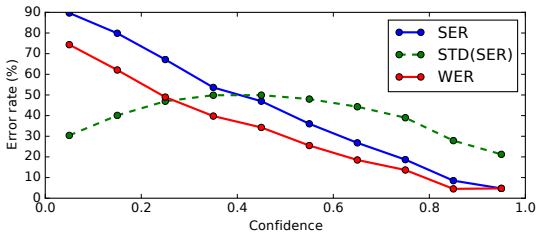

Figure 2: WER, SER, and the standard error of SER, computed over 10 uniformly spaced confidence bins.

error curve derives from the fact that SER of an utterance is a binary variable. Hence each bin i can be modeled as a Bernoulli process with a success probability $p_i = \overline{SER}_i$ and a sample variance $s_i^2 = p_i(1 - p_i)$.

3. Analyzing evaluation metric reliability

In typical real-life scenarios, we cannot afford to label all the data that is available. Instead, we need to work with a subset, a sample of the data, which we call a *dataset*. The random selection of the dataset utterances affects the evaluation results computed over it. A metric extracted over the dataset is in fact a random variable, and may deviate from its true value (that calculated over the whole population). In this section, we investigate ways to formalize this sampling effect and how to analyze the precision of a metric.

3.1. Variance of mean estimator

Many metrics of interest can be obtained as an average over individual data instances. The *sample mean* $\bar{x}$ is an unbiased estimator of the true *population mean* $\mathrm{E}(X)$:

$$\bar{x} = \frac{1}{N} \sum_{j=1}^{N} x^j, \tag{1}$$

where x^j are data instances, randomly sampled from the population, and N is the sample (dataset) size. In ASR tasks, a data instance x^j could represent e.g. the SER of an utterance j. Due to the random selection of x^j, sample mean is a random variable, with variance inversely proportional to the sample size:

$$\mathrm{Var}(\bar{x}) = \frac{s^2}{N}, \tag{2}$$

where s^2 is the sample variance of the data instances x^j.

3.2. Bootstrapping

Eq. (2) can be used to study the variance of evaluation metrics such as SER. However, the distribution of the deviations with a finite dataset size remains unknown. Furthermore, metrics such as WER cannot be computed as an average of the instance values, in which case (2) cannot be directly applied. To allow flexible analysis of sampling effects, we used *Bootstrapping* [8, 9]. It resamples the original dataset with replacement to form "samples of datasets", with desired dataset sizes. These new datasets can be used to produce empirical distributions and statistics of the evaluation metrics. The advantage of using bootstrapping is that it does not set strict assumptions on the distribution of the data, as long as the population variance of the metric in interest is finite. Using bootstrapping we can also derive the 95% quantiles of the error estimates, which we will use to quantify the precision of the estimates.

4. Stratified sampling

Stratified sampling [1, 2] is a well-known method to ensure sufficient coverage of different types of events in the sample. The idea is to partition the population into groups called *strata*, and randomly sample from each group (*stratum*) separately. The allocation of the samples into strata can be varied, and it is taken into account to normalize the final statistics computed from the sample, so that unbiased estimates are produced. The method can be used to minimize estimator variances of sample statistics, or to produce representative samples out of various subsets of the population while maintaining an unbiased estimate of the overall sample mean.

4.1. Formulation

In stratified sampling, the mean of a random variable is computed as a weighted average of the means over each stratum:

$$\bar{x}_s = \sum_{i=1}^{m} \frac{N_i}{N} \bar{x}_i = \sum_{i=1}^{m} \frac{N_i}{N} \frac{\sum_{j=1}^{n_i} x_i^j}{n_i}, \tag{3}$$

where N_i is the number of instances for stratum i in the whole population, $N = \sum N_i$, n_i is the number of instances randomly drawn for stratum i, m is the number of strata, and x_i^j is a data instance in stratum i. The variance of $\bar{x}_s$ acquires the following form:

$$\mathrm{Var}(\bar{x}_s) = \sum_i \frac{N_i^2}{N^2} \mathrm{Var}(\bar{x}_i) = \sum_i \frac{N_i^2}{N^2} \frac{s_i^2}{n_i}, \tag{4}$$

where the last equality uses (2) with s_i^2 as the sample variance of i^{th} stratum.

4.2. Defining the strata

The prerequisite for applying stratified sampling is that all the data can be uniquely and exhaustively assigned to the defined strata. The strata can be defined based on category variables known before the sampling. This is especially useful if one wishes to analyze the subsets defined by the category, as stratified sampling can improve the precision of the subset statistics. Another option is to use a continuous variable and discretize it into different bins defined as the strata. This is the approach adopted in this paper.

Eq. (4) suggests that by choosing the strata in such a way that s_i^2 are smaller than the overall s^2 (that is, x_i^j of different

strata are not identically distributed), the estimator variance becomes smaller. In addition, by altering the strata allocation n_i, we can control the sample mean variances and reduce the estimator variance further.

With statistical classifiers, a useful way to partition the data is to use the classifier score as the stratification variable [3]. By defining the strata as non-overlapping ranges of the score, the model itself defines the partition in an unsupervised manner. As mentioned in Section 2.2, the model confidence values of our ASR system correlate well with the error metrics of the predictions. Hence different confidence ranges have different distributions of the error metrics, which makes them a good choice as the strata.

When defining the partitioning based on a continuous variable, one has the freedom to choose the desired number of strata, as well as ranges of values corresponding to each strata. The strata are defined prior to sampling, so we do not have access to the sample distributions, although we may have historical data to derive expectations from. The strata should be defined such that we can trust the estimates of the strata frequencies N_i/N.

A simple way for data partitioning with a continuous stratification variable is to use *uniform ranges* over the values of the stratification variable. However, if the stratification variable is not uniformly distributed, it might be better to use *variable ranges*, in such a way that approximately the same number of instances are expected to be observed in each stratum. We will investigate both of these strategies in this paper.

4.3. Sample allocation

To apply stratified sampling, we need to define the *sample allocation* n_i, the number of instances drawn for each stratum. In *proportional allocation* strategy, n_i is defined to be proportional to the probability of observing a data instance in the stratum:

$$\frac{n_i}{n} = \frac{N_i}{N} \qquad \text{for } i = 1 \dots m, \tag{5}$$

$$\sum_i n_i = n. \tag{6}$$

However, n_i doesn't have to be strictly proportional to real frequencies of the strata. In fact, it is beneficial to set sample allocations n_i such that the variance of the sample mean estimate is minimized. This strategy is called *optimal allocation*.

Optimal allocation uses the sample variances of each stratum to guide the allocation. It can be derived from (4) using e.g. Lagrange multipliers [10]. This leads to so called Neyman allocation, which states that the sample size for stratum i is

$$n_i = n \frac{N_i s_i}{\sum_j N_j s_j}. \tag{7}$$

Optimal allocation assigns more samples to strata with large sample variances, hence minimizing their contribution to the total variance of the mean estimate. If the sample variances are not known prior to sampling, sampling and labeling can be done incrementally [3]. When considering optimal allocation, the number of strata poses a trade-off: large number of strata may provide lower variances, but the less samples there are for each stratum, the harder it is to estimate the variance within each strata, leading to deviations from the optimal strategy [11].

WER cannot be computed as a weighted sum of the stratum estimates. Instead, stratum errors and reference lengths need to be summed separately:

$$\bar{W}_s = \frac{\sum \frac{N_i}{N} \bar{e}_i}{\sum \frac{N_i}{N} \bar{r}_i}, \tag{8}$$

where $\bar{e}_i$ and $\bar{r}_i$ are the sums of errors and reference lengths in stratum i, respectively. Due to this formulation, using (7) directly is not possible. Instead, we propose a novel strategy which provides approximately optimal sample allocations for WER. We start by considering the 1^{st} order Taylor approximation of (8) and derive the total variance of WER as

$$\mathrm{Var}(\bar{W}_s) \approx \sum_i \left[\left(\frac{\partial W}{\partial \bar{e}_i} \right)^2 \mathrm{Var}(\bar{e}_i) + \left(\frac{\partial W}{\partial \bar{r}_i} \right)^2 \mathrm{Var}(\bar{r}_i) + 2 \frac{\partial W}{\partial \bar{e}_i} \frac{\partial W}{\partial \bar{r}_i} \mathrm{Cov}(\bar{e}_i, \bar{r}_i) \right]. \tag{9}$$

We will use the following partial derivatives:

$$\frac{\partial W}{\partial \bar{e}_i} = \frac{N_i}{N r_{avg}}, \tag{10}$$

$$\frac{\partial W}{\partial \bar{r}_i} = \frac{-N_i e_{avg}}{N r_{avg}^2}, \tag{11}$$

where e_{avg} and r_{avg} are the average number of word-level errors and number of reference words in an utterance over the population, respectively. After substituting the variances in (9) with sample estimates (2), we can use Lagrange multipliers to solve the sample allocations n_i which minimize (9), subject to constraint (6):

$$n_i = \frac{k N_i}{r_{avg}^2} \sqrt{r_{avg}^2 s_{\bar{e}_i}^2 + e_{avg} s_{\bar{r}_i}^2 - 2 r_{avg} e_{avg} s_{\bar{e}_i, \bar{r}_i}}, \tag{12}$$

where k is a constant such that the constraint (6) is satisfied, and $s_{\bar{e}_i, \bar{r}_i}$ is the sample covariance between $\bar{e}_i$ and $\bar{r}_i$.

5. Experiments

In this section we provide experimental evidence on the advantage of using stratified sampling for evaluating ASR systems. Throughout the analysis in this section, the deviations of evaluation metrics are measured as symmetric 95% quantiles, expressed relative to the absolute error metric.

5.1. Deviation of SER with different sampling methods

Figure 3 illustrates the effect of the sample size, when using random sampling or stratified sampling with 10 uniformly distributed confidence bins, using proportional or optimal sample allocation. As expected, increasing the sample size increases the precision of mean estimates. Comparing the quantiles of random and stratified sampling shows that for the same deviation, stratified sampling requires about 30% fewer utterances. Optimal allocation slightly but consistently outperforms proportional allocation strategy.

Table 1 shows the 95% quantiles of SER estimates over evaluation sets of 10000 utterances using different stratification strategies. We can observe that already the simplest stratification scheme, uniform confidence ranges and proportional allocation, improves the precision significantly over the random sampling. As in Figure 3, optimal allocation always outperforms the proportional one. Variable confidence ranges do not seem to be beneficial, and may even degrade the precision compared to uniform ranges. This is likely to be a property of the

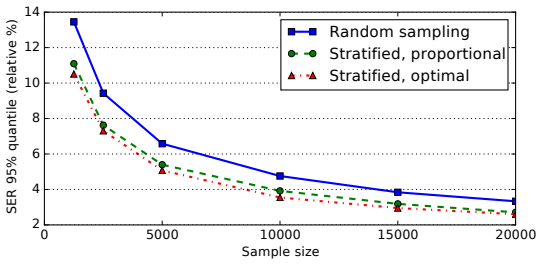

Figure 3: The effect of sample size on the SER deviation, measured as a symmetric 95% quantile relative to the observed SER value. Random sampling shows 28–34% wider quantiles than stratified sampling with optimal allocation (10 uniform bins).

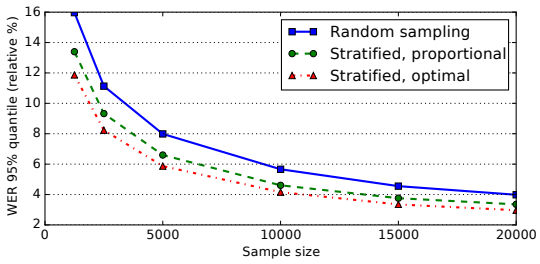

Figure 4: Deviation of WER metrics as relative symmetric 95% quantiles. Stratification used 20 bins with uniform ranges.

Table 1: Relative SER deviations as 95% quantiles when using different stratification strategies: Number of strata, definition of strata (as uniform or variable ranges), and different allocation strategies (proportional or optimal). The number of samples was kept fixed at 10000 utterances. The 95% quantile for random sampling in this case is $\pm 4.7\%$.

# of	Uniform bins		Variable bins	
strata	Proport.	Optimal	Proport.	Optimal
5	3.9%	3.7%	4.0%	3.8%
10	3.9%	3.6%	3.9%	3.7%
20	3.7%	**3.5%**	3.9%	3.6%
30	3.8%	3.6%	3.9%	3.7%
40	3.8%	**3.5%**	3.9%	3.7%

particular task and the model at hand. With the very good correlation between the confidence values and the error rates provided by the model, uniform confidence ranges match better with the differences in the distributions of errors. Variable confidence range strategy, on the other hand, results in too much pooling for the less frequent low-confidence utterances.

Increasing the number of strata is seen to be beneficial up to 20 strata. Using more strata does not provide additional gains. As lower number of strata is preferred, due to more robust estimates of strata frequencies and error variances, 20 strata with uniform confidence bins is considered to be optimal for this task.

5.2. Stratification using acoustic measures

Ideally stratification could be defined independently of the model, such that model updates would not change the optimal sample allocation. To achieve this, we tried using purely acoustic measures instead of model confidences to define the stratification. Low signal-to-noise ratio (SNR) and extreme fundamental frequencies F_0 are known to degrade recognition accuracies, so both of them were tested as stratification variables. Tests were run with 10 and 20 strata, using both uniform and variable ranges, with dataset size of 10000 utterances. In addition, SNR and F_0 were tested in combination to define 25 strata as a uniform 5x5 grid. In all the cases, utterances without speech contents were assigned to a dedicated stratum, as estimating SNR or F_0 for them would not be possible. F_0 contours were extracted using the Summation of the Residual Harmonics (SRH) algorithm [12]. SRH was also used for voice activity detection, in a method similar to [13, 14].

Despite the various configurations, none of the tested acoustic-based stratification methods provided gains over random sampling. Although low/extreme values for SNR/F_0 indicate higher error rates, majority of the errors in the evaluation set occur in utterances with typical values of the two measures. Using model confidences is therefore a superior method for stratification. Furthermore, gradual model updates may not pose such a problem, as sample allocation optimum has been shown to be flat in the sense that small deviations in the allocations cause only small variations in the variance [10].

5.3. Measuring WER with stratified sampling

In Section 4.3 we introduced an approximation for optimal sample allocation applicable for WER. Figure 4 shows WER 95% symmetric quantiles for random sampling, proportional allocation, and the optimal WER allocation, computed with 20 uniform bins. Stratified sampling and optimal allocation outperforms other two methods. As an alternative, one could consider using SER as a proxy to define the sample allocation with (7), even though WER is computed over the strata. However, our experiments showed that this approach produces 6–8% larger WER quantiles, compared to using (12).

6. Conclusions

This paper has demonstrated how stratified sampling can be used to optimize the evaluation of ASR, by reducing the deviations observed in the error metrics such as WER or SER. Stratified sampling with model confidences as stratification variable allowed using about 30% fewer utterances than random sampling would require for a given precision of SER. This can be directly translated into reducing the labeling cost of the evaluation sets.

Various configurations for stratified sampling were experimentally compared. Optimal allocation was shown to provide the best precision, provided that we have prior knowledge on the distribution of the error metrics over confidence ranges, so that the strata can be define sensibly. Defining the strata as uniformly distributed ranges of the model confidence scores was shown to work well when confidences correlated well with the error metrics. Finally, we introduced a novel method for defining approximately optimal sample allocations for WER evaluations, and showed that it outperformed other allocation methods.

7. References

[1] J. Neyman, "On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection," *Journal of the Royal Statistical Society*, vol. 97, no. 4, pp. 558–625, 1934.

[2] W. G. Cochran, *Sampling Techniques*, 3rd ed., ser. Wiley series in probability and mathematical statistics: Applied probability and statistics. Wiley, 1977.

[3] P. N. Bennett and V. R. Carvalho, "Online stratified sampling: Evaluating classifiers at web-scale," in *Proc. 19th ACM International Conference on Information and Knowledge Management*, ser. CIKM '10, 2010, pp. 1581–1584.

[4] G. Druck and A. McCallum, "Toward interactive training and evaluation," in *Proc. 20th ACM International Conference on Information and Knowledge Management*, ser. CIKM '11, 2011, pp. 947–956.

[5] H. Jiang, "Confidence measures for speech recognition: A survey," *Speech Communication*, vol. 45, pp. 455–470, 2005.

[6] D. Yu, J. Li, and L. Deng, "Calibration of confidence measures in speech recognition," *IEEE Transactions on Audio, Speech and Language*, vol. 19, no. 8, pp. 2461–2473, 2011.

[7] G. Evermann and P. Woodland, "Posterior probability decoding, confidence estimation and system combination," in *Proc. Speech Transcription Workshop*, vol. 27, 2000.

[8] B. Efron, "Bootstrap methods: Another look at the jackknife," *The Annals of Statistics*, vol. 7, no. 1, pp. 1–26, 1979.

[9] P. Koehn, "Statistical significance tests for machine translation evaluation," in *Proc. EMNLP*, 2004, pp. 388–395.

[10] J. B. Kadane, "Optimal dynamic sample allocation among strata," *Journal of Official Statistics*, vol. 21, no. 4, pp. 531–541, 2005.

[11] A. Carpentier and R. Munos, *Algorithmic Learning Theory - 23rd International Conference, ALT 2012, Lyon, France, October 29-31, 2012. Proceedings*. Springer Berlin Heidelberg, 2012, ch. Minimax Number of Strata for Online Stratified Sampling Given Noisy Samples, pp. 229–244.

[12] T. Drugman and A. Alwan, "Joint robust voicing detection and pitch estimation based on residual harmonics," in *Proc. Interspeech*, 2011, pp. 1973–1976.

[13] S. O. Sadjadi and J. H. L. Hansen, "Unsupervised speech activity detection using voicing measures and perceptual spectral flux," *IEEE Signal Processing Letters*, vol. 20, no. 3, pp. 197–200, 2013.

[14] T. Drugman, Y. Stylianou, Y. Kida, and M. Akamine, "Voice activity detection: Merging source and filter-based information," *IEEE Signal Processing Letters*, vol. 23, no. 2, pp. 252–256, 2016.