Unsupervised Testing Strategies for ASR

Brian Strope, Doug Beeferman, Alexander Gruenstein, Xin Lei

Google, Inc.

bps, dougb, alexgru, xinlei @google.com

Abstract

This paper describes unsupervised strategies for estimating relative accuracy differences between acoustic models or language models used for automatic speech recognition. To test acoustic models, the approach extends ideas used for unsupervised discriminative training to include a more explicit validation on held out data. To test language models, we use a dual interpretation of the same process, this time allowing us to measure differences by exploiting expected 'truth gradients' between strong and weak acoustic models. The paper shows correlations between supervised and unsupervised measures across a range of acoustic model and language model variations. We also use unsupervised tests to assess the non-stationary nature of mobile speech input.

Index Terms: speech recognition, unsupervised testing, non-stationary distributions

1. Introduction

Current commercial speech recognition systems can use years of unsupervised data to train relatively large, discriminatively optimized, acoustic models (AM). Similarly, web-scale text corpora for estimating language models (LM) are often available online, and unsupervised recognition results themselves can provide an additional source of LM training data.

Since there is no human transcription in any of these steps, the remaining use for manual human transcription is for generating test sets, as a final sanity check for validating system parameters and models. In this paper, we augment that strategy with unsupervised evaluations and begin the discussion of whether eventually we might be able to get rid of the need for any explicit human transcription.

The motivation for human transcription for testing is obvious. Despite steady advances and relative commercial successes, it is generally accepted that humans are much more accurate transcribers than automatic speech recognition systems [1]. While there are a few notable exceptions where machines were more accurate than humans [2], human transcription accuracy is so much better, we use it unquiestioningly as our best approximate for absolute truth.

But there are equally obvious disadvantages to relying on human transcription. While it may feel premature, accepting human performance as absolute truth imposes an upper bound on accuracy. The absolute truth is not absolute, and so we'll eventually have to figure out how to beat it. In fact with our current processes and tasks, below, we show that human transcribers can be only comparable in accuracy to current ASR systems. Absolute truth is already a problem. In response, we are improving transcription processes, but also considering unsupervised ways to augment traditional testing.

Another obvious disadvantage of human transcription is that the tests themselves have to be limited in size and type. Even in a commercially successful research lab, getting extensive tests across every combination of speaker and channel type, recognition context, language, and time period is prohibitive. But a detailed characterization of those types of variations could help prioritize efforts. Similarly when tests are unsupervised, it is easier to update development and evaluation sets to avoid problems related to stale, over-fit tests.

This is mostly an empirical paper. The next section describes some of the experiments we ran trying to assess our existing human transcription accuracy. Then we describe the generalizations of unsupervised discriminative training that enable a new evaluation strategy. Next the paper includes evaluations that show correlations between supervised and unsupervised tests, and concludes with unsupervised tests that start to characterize the non-stationary distribution of spoken data coming through Google mobile applications.

2. Problems with human transcriptions

Recent efforts have begun to consider human transcription accuracy in the context of increased efficiency. These studies have generally shown that depending on the amount of effort, and the task, individual word error rates can vary from 2-15% [3, 4]. Efficiency pressures on human transcription can lead to transcription noise and bias.

2.1. Early experiments

Over the last few years we have seen several simple experiments not work: we have added matched data to our language models and seen error rates get worse; we have added unsupervised acoustic modeling data matched to a new fielded acoustic condition, and seen the error rates on new matched tests go up, but surprisingly, error rates on an old test, with slightly mismatched conditions, go down.

For each of these, after tediously examining errors, we found the problem was that we typically "seed" our transcription process with the recognition result from the field. Mostly as a matter of expedience; it is easier for the transcriber to hit return than to type "home depot in palo alto california" yet again, and it can improve reliability since retyping can be error prone. But the power of the suggested transcription is also enough to bias the transcribers into rubber-stamping some of the fielded recognition results. When the transcriber rubber-stamps an error we potentially get penalized twice. The baseline gets credit where it should not, and a new system that corrects that error is falsely penalized for adding an error.

The surprising improvement noted on the older, slightly mis-matched test happened because the transcriptions for the older test were seeded with transcriptions from an older system, decorrelating some of the transcription bias with the current baseline. In this case, transcription bias toward the baseline model was a bigger effect than the change in acoustics.

2.2. Multiple attempts

To measure the human transcription accuracy more directly we started sending the same data for multiple attempts at human transcription, and we intentionally reduced the quality of our starting seeds to move any bias away from our best systems. For one test we sent 200K Voice Search utterances to be transcribed twice. Ignoring trivial differences like spaces, apostrophes, function words, and others, half of the transcripts agreed, which implies a sentence transcription accuracy of 71%, assuming independence of the attempts.

Similarly when we sent the remaining 100K utterances, where transcriptions did not agree, back for two more attempts, we were still left with about 10% of the original set with 4 distinct human transcriptions. Again assuming independence, 10% disagreement in 4 attempts is consistent with 68% accuracy for each attempt. But we believe our system has a sentence accuracy higher than 70%.

Looking through the errors many of the problems are related to cultural references, popular names, and businesses that are not obvious to everyone. The cultural and geographic requirements of the voice search task may be unusually difficult. It combines short utterances and wide open semantic contexts to generate surprisingly unfamiliar sounding speech. Finding ways to bring the correct cultural context to the transcriber is another obvious path to pursue.

3. Generalizing unsupervised discriminative training

While some published results considered unsupervised maximum likelihood estimation of model parameters [5], many systems use unsupervised discriminative optimization, directly using recognizer output as input [6]. Cynically we might ask what we are learning if we are using the recognition result as truth for discriminatively optimizing its parameters. It is hard to imagine that we can fix the errors it makes, when we use the model to generate truth.

But when we look into the details of commonly used discriminative training techniques based on maximum mutual information, we see that the LM used to generate competing hypotheses is not the same LM used to generate truth. To improve the generalization of discriminative training, we use a unigram to describe the space of potential errors [7], but a trigram or higher to give us transcription truth with unsupervised training.

One interpretation of unsupervised discriminative training for acoustic models is that we are using the difference between a weak unigram and a relatively stronger trigram to give us a known improvement in relative truth. We do not know that the strong-LM (trigram) result is absolutely correct, we only know that it is better than the result with the weak LM (unigram). When there is a difference, if we can move toward the results of the strong-LM system by changing acoustic model parameters, then we are building a more accurate AM, that also helps with the final system using a stronger LM. With this interpretation, the AM learns from the 'truth gradient' between the strong and weak LMs.

3.1. Unsupervised AM testing

Extending unsupervised discriminative AM training to unsupervised AM testing involves retesting the criterion used during training in a new test context. More prescriptively, we sample a new set of live data from production logs, and take the recognition result from the fielded system using a strong AM and LM as assumed truth. Then we re-recognize the same data using multiple strong acoustic models and a weak LM. If one of the systems using a weak LM can better approximate the system using a strong LM, then at a minimum, we can say that it is doing a better job of generalizing our training criteria to new data. More directly, we have evidence that one of the strong acoustic models could be more accurate than the rest.

For scoring we are assuming truth from the fielded system, not a human transcriber. Therefore, when reporting unsupervised testing results, we count traditional word error rates, but because there is no human transcription, we report it as a word difference rate (WDR), to highlight that, for example, in the case of unsupervised AM tests, it is the word differences between the systems with the strong and weak LM.

3.2. Unsupervised LM testing

To use the same strategy for LM testing we reverse the roles of the AM and the LM. For better generalization of discriminative AM testing, we used a weak LM to generate more competing alternates. That establishes a truth gradient that generally changes around 1/3 of the words. The dual for LM testing is to use a weak AM instead. To get a truth gradient of a similar magnitude with our systems, we backed off to a context-dependent acoustic model that uses around 1/10th the number of parameters of our strong models, and only uses maximum likelihood training.

Then as above, we test with multiple strong LMs and assume that the LM that can move the results of the system using the weak AM closest to the results of the production system (with the strong AM), is the most accurate LM. With unsupervised LM testing we again report WDR and not WER, where the magnitude of the difference is now from the difference between the strong AM and the weak AM.

3.3. Relative measures

In this paper we are ignoring the harder problem of measuring absolute accuracy. Instead we focus on relative differences between different acoustic or language models. Others have predicted absolute error measures using statistics from the training set as represented in the final acoustic models [8], without looking at testing data. But here we are interested in estimating relative performance across production data that was unseen during training. Our goal is to assess whether new models or new approaches are helping on new data, and whether the data might be changing from the distributions used during training.

4. Correlating supervised and unsupervised measures

First we show that the performance on unsupervised offline tests for the AM and for the LM correlate with more traditional supervised tests. Our production data started with primarily Voice Search queries intended for google.com, but over time has included increasing amounts of general Voice Input traffic which includes a large fraction of short person-to-person messages. To start the analyses, we consider these data streams separately.

For Voice Search, our traditional supervised test is built from the 200K utterance set that we sent for multiple transcriptions. For this test we exclude the 10% of the utterances where we got 4 distinct human transcriptions and sample a test set randomly from the remaining 90%. Similarly for the supervised

Voice Input test, we sent utterances twice and selected from the utterances with at least 80% agreement between human transcriptions. On the utterances where not all the words agreed, we randomly chose one of the human transcriptions as truth. This led to a test that excluded about 28% of the utterances.

Both of these supervised tests are biased in that they only include the utterances that we could reliably transcribe. The Voice Search test has 27K utterances and 87K words. The Voice Input test has 49K utterances and 320K words.

For the first unsupervised tests here, we sampled production logs for a single day of traffic. We found the median recognizer confidence for each task and then randomly selected a few hundred thousand utterances that were above median confidence for each task. For all unsupervised experiments we used the recognition results from the field as truth.

Our recognition configuration for both systems is fairly standard and described in the literature. Specifically we use a PLP front-end [11] together with LDA and STC [12], and optimize our acoustic models using BMMI [13] on mostly unsupervised data mixed from both tasks. Our language models are n-grams, with Katz interpolation and entropy pruning, and the fielded Voice Input system also includes dynamic interpolation [14]. The Voice Search system used trigrams and the Voice Input system included 4-grams.

4.1. AM experiments

The AM experiments use a weak LM (in this case a unigram) for each task estimated from the few hundred thousand high confidence utterances sampled for that day's test. All the utterances in the test were also used to train the LM, so there is no OOV. This step is consistent with the matched unigram we train for discriminative acoustic model training. For Voice Search, the resulting unigram had 17K words, and for Voice Input there were 18K unique words.

The acoustic models we tested here were trained using 11M (mostly unsupervised) utterances from a mix of both tasks. The parameter we vary for these experiments is the size of the acoustic models. We use the same decision tree and context state definitions for all models, but we vary the number of Gaussians assigned to each state. Each model is trained with the same number of iterations through all the data. The final model sizes range from 100K to 1M Gaussians. Decoder parameters are set in production mode, which generally means we lose around 0.5% absolute from the best possible accuracy to have faster than real-time search.

# Gauss	Sup VS	Unsup VS	Sup VI	Unsup VI
100K	16.0	36.0	14.5	24.8
200K	15.3	34.4	13.6	22.8
340K	14.6	33.9	13.4	22.7
500K	14.3	33.3	13.2	22.3
1M	13.9	33.0	12.9	21.8

Table 1: *WER in % on supervised (Sup) and WDR in % on unsupervised (Unsup) AM tests for Voice Search (VS) and Voice Input (VI).*

4.2. LM experiments

For the LM experiments we vary the number of n-grams used for the Voice Input task from around 2M to 30M by varying our final entropy pruning threshold. Unlike the production system used to generate truth for the unsupervised tests, for these tests the LM is a static n-gram.

We show results with two different weak acoustic models (A/B). Condition A is a context-dependent model estimated using maximum likelihood criteria with 2 Gaussians per state for a total of 16K Gaussians. Condition B uses a similar model with a variable number of Gaussians across model states, and a total of 40K Gaussians. On supervised tests, these weak acoustic models have around two to three times the error rates of final strong production models.

n-grams	Sup PPL	Sup WER	Unsup A/B WDR
1.9M	109	15.2	38.1/25.9
3.8M	98	14.4	36.8/24.5
7.6M	92	14.1	36.0/23.8
15M	87	13.9	35.5/23.2
30M	85	13.7	35.1/22.8

Table 2: *Comparing supervised (Sup) and unsupervised (Unsup) LM tests for Voice Input. WER/WDR are in %, PPL is perplexity. Unsup A and B are for different sized AMs.*

The relative improvement in both AM and LM experiments is consistently around 10% for a 10x increase in model size. Correlations between supervised and unsupervised tests range between 0.98 and 0.99.

5. Additional experiments

Varying model size is a controlled way to generate accuracy differences. Here we include additional unsupervised measurements that show expected differences in the context of other AM and LM modeling efforts.

5.1. CMLLR

To evaluate an implementation of constrained maximum likelihood linear regression [9] for adaptation, we started by testing with read speech corpora from several data collections [10] used to initialize acoustic models in a new context. With a large and regular amounts of acoustic data per speaker, we see the typical improvements of 6-10% relative, over a matched discriminative baseline.

To estimate the accuracy impact of CMLLR on the production system, (where the actual distributions of amount of data per user is not imposed by the strict specifications of a data collection) we used unsupervised testing. Here we sampled all personalized users over a 30 day period, and measured the change in WDR with a weak LM and either the production AM or the production AM with CMLLR. Further we break the differences in WDR down by the amount of data available for each speaker.

# Utts	No Adapt	Adapt
1-20	25.7	25.4
20-50	26.6	25.6
50-100	25.8	24.6
100-200	23.5	22.5

Table 3: *WDR in % on adaptation tests. Input is binned by the number of utterances for a given user.*

From the table, it is clear that we are seeing a similar relative difference as we saw with more traditional read speech tests, and we are further able to characterize the expected satu-

ration of the relatively small number of parameters in CMLLR after around 20 voice input utterances.

5.2. LM update

At one point we updated our language model to include a rescoring pass more explicitly matched to recent Voice Search queries. By testing this update with recent unsupervised tests we are able to show the expected win on new voice search type utterances.

# Model Config	Sup VS	Unsup VS
Original	14.6	30.0
Updated	14.6	28.6

Table 4: *WER in % on supervised (Sup) and WDR in % unsupervised (Unsup) LM tests for Voice Search.*

One interpretation of these results is that we are updating the LM to better represent the recent query data which itself is better matched to the recent unsupervised test. It also suggests that the distribution of our data might be moving.

5.3. Estimating non-stationary distributions

Finally we ran two sweeps of AM tests to estimate how stationary the acoustics for our system have been over the last 14 months. The first system is trained using the Voice Search supervised data available at the beginning of the 14 months, and the second uses only unsupervised data sampled from the last 3 months. Therefore, one model represents our initial estimate of the distribution, and the other approximates a most recent distribution. Both systems use around 350K gaussians. To evaluate the AM performance, we use a weak LM estimated from a year's worth of production data.

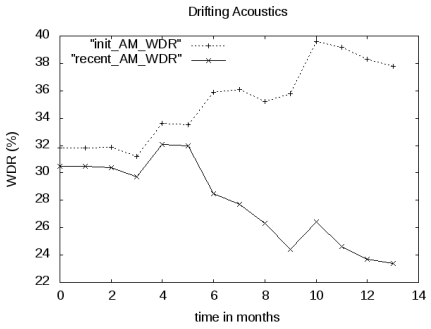

Figure 1: *Change in WDR over time with two different AMs.*

Both lines show that the distribution of the data has shifted away from the original supervised data, and toward the recent unsupervised data. Additional unsupervised tests will illuminate the causes of this change in more detail. We currently suspect an increase in the fraction of voice input recognition, but it is already obvious that the distribution of the acoustics for this data is changing. The plot also suggests that with a single AM the change of WDR across conditions may also be informative.

Note that since we are generalizing from the same criteria we used for AM training, and we are getting rid of some of the necessity of human transcription, we are concerned about converging away from reality. The ground is a little firmer for the LM side, since our current LM processes are in fact not yet learning from AM truth gradients the same way our unsupervised AM training learns from LM truth gradients. From the AM side, our current unsupervised tests are simply checking whether the training optimizations extend to unseen data. Pragmatically, because it is unsupervised we also have the opportunity to test that generalization with a range of weak LMs and with a range of input data, and thereby to increase our confidence in the generalization. Moreover, reducing the accuracy improvement provided by a strong LM seems like a safe requirement to impose on AM training. But from an experiment perspective, we have to remember what gradient we are exploiting and not cheat. In other words, augmenting the AM with features directly related to the strong LM would not lead to improvements. We also monitor coarse signals related to application use (counts of user actions in response to recognition results) to give us additional complimentary evidence of successful generalization.

6. Conclusions

This paper extends unsupervised discriminative training to an unsupervised testing strategy suitable for evaluating AM and LM changes. We show strong correlations with traditional testing strategies when we change AM or LM model size. We also show expected gains on unsupervised measures with other types of AM and LM changes, and use the unsupervised measures to begin to characterize the stationarity of the input data to Google mobile. Together with unsupervised training, unsupervised testing enables development paths that no longer impose human performance as the upper bound for accuracy.

7. References

[1] R. Lippmann, "Speech recognition by machines and humans," Speech Communication, July 1997.

[2] T. Kristjansson, J. Hershey, P. Olsen, S. Rennie, R. Gopinath, "Super-Human Multi-Talker Speech Recognition: The IBM 2006 Speech Separation Challenge System," Proc. ICSLP, 2006.

[3] S. Novotney, C. Callison-Burch, "Cheap, Fast and Good Enough: Automatic Speech Recognition with Non-Expert Transcription," Proc. NAACL, 2010.

[4] A. Gruenstein, I. McGraw, A. Sutherland, "A Self-Transcribing Speech Corpus: Collecting Continuous Speech with an Online Educational Game," Proc. SLaTE, 2009.

[5] J. Ma, R. Schwartz, "Unsupervised versus supervised training of acoustic models," Proc. ICSLP, 2008.

[6] L. Wang, M. Gales, P. Woodland, "Unsupervised Training for Mandarin Broadcast News Conversation Transcription," Proc ICASSP, 2007.

[7] P.C. Woodland, D. Povey, "Large scale discriminative training of hidden Markov models for speech recognition," Comp. Speech & Lang., Jan. 2002.

[8] Y. Deng, M. Mahajan, A. Acero, "Estimating Speech Recognition Error Rate without Acoustic Test Data," Proc. Eurospeech, 2003.

[9] M. J. F. Gales, "Maximum likelihood linear transformations for HMM-based speech recognition," Comp. Speech & Lang., Vol 12.2 1998.

[10] T. Hughes, K. Nakajima, L. Ha, A. Vasu, P. Moreno, M. LeBeau, "Building transcribed speech corpora quickly and cheaply for many languages," Proc ICSLP, 2010.

[11] H. Hermansky, "Perceptual linear predictive (PLP) analysis of speech," JASA, v87.4, 1990.

[12] M. Gales, "Semi-Tied Covariance Matrices for Hidden Markov Models," Proc. IEEE Trans. SAP, May 2000.

[13] D. Povey, D. Kanevsky, B. Kingsbury, B. Ramabhadran, G. Saon, K. Visweswariah, "Boosted MMI for model and feature-space discriminative training," Proc. ICASSP, 2008.

[14] B. Ballinger, C. Allauzen, A. Gruenstein, J. Schalkwyk, "On-Demand Language Model Interpolation for Mobile Speech Input," Proc. ICSLP, 2010.