A big data approach to acoustic model training corpus selection

Olga Kapralova, John Alex, Eugene Weinstein, Pedro Moreno, Olivier Siohan

Google Inc.
76 Ninth Avenue, New York, NY 10011

Abstract

Deep neural networks (DNNs) have recently become the state of the art technology in speech recognition systems. In this paper we propose a new approach to constructing large high quality unsupervised sets to train DNN models for large vocabulary speech recognition. The core of our technique consists of two steps. We first redecode speech logged by our production recognizer with a very accurate (and hence too slow for real-time usage) set of speech models to improve the quality of ground truth transcripts used for training alignments. Using confidence scores, transcript length and transcript flattening heuristics designed to cull salient utterances from three decades of speech per language, we then carefully select training data sets consisting of up to 15K hours of speech to be used to train acoustic models without any reliance on manual transcription. We show that this approach yields models with approximately 18K context dependent states that achieve 10% relative improvement in large vocabulary dictation and voice-search systems for Brazilian Portuguese, French, Italian and Russian languages.

Index Terms: large unsupervised training sets, data selection, deep neural networks, acoustic modeling

1. Introduction

Automatic speech recognition (ASR) systems have become a popular human-computer interaction modality and have been deployed as an input method in many successful commercial products [1]. The quality of available ASR systems has seen significant improvement over the last decade for various small and large vocabulary recognition problems [2]. In particular, one of the main driving forces behind the most recent improvements are deep neural networks (DNN), which have become the state of the art technology for acoustic modeling [2].

It has been observed [3] that the performance of DNN models improves with the number of context-dependent states and hidden layers used. The high dimensionality of the model, however, requires large volumes of high quality data to be used for training. Traditionally, acoustic models are constructed from supervised data manually transcribed by humans. Unfortunately, manual transcription of large data sets is expensive and scales poorly for systems that support a large number of languages, such as Google speech products. Rather than relying on small supervised training sets, one can instead exploit the large amount of audio data that is processed by systems like Voice Search, together with semi-supervised techniques to construct large training sets with automatically derived transcripts [4, 5, 6, 7, 8].

In this paper we describe Google's approach to constructing speech corpora and models in a fully unsupervised manner. Our process is based on redecoding a massive amount of anonymized audio logs from Google's speech products using a high accuracy, non real-time system, followed by a series of heuristics aiming at discarding incorrectly recognized speech recordings. Such an automated procedure enables the collection of training sets an order of magnitude larger than what we can typically obtain by human transcription. Moreover, our use of a large-scale distributed computation platform enables training sets to be produced in a matter of days, as compared to the months required to produce manually transcribed training sets with sufficient quality. Automatic generation of training sets also enables us to easily track the continuous changes in the fleet of mobile and desktop devices that use Google Voice Search and Voice Input systems. For example, during 2013 more than 1.5 million new Android devices were activated per day, and new models with different hardware and microphones arrive on the market daily.

We present recognition results on the Voice Search and Voice Input tasks for Brazilian Portuguese, French, Italian and Russian languages, based on training corpora automatically generated using the proposed approach.

The paper is organized as follows. In Section 2 we describe our methodology for selecting the training corpus from logs of our production recognizer. Section 3 describes our training system setup. Finally, in Section 4 we report on the accuracy of the new production system obtained by our method.

2. Methodology

High-quality acoustic models rely on the availability of large and reliably transcribed training sets that match the underlying distribution of speech in different acoustic environments. It is also imperative that the training data represent the demographic variety of the potential speakers well. Such audio data sets can be constructed from the logs of a deployed ASR system.

In this section we describe our big data driven approach for selecting a training corpus from these logs. We first describe the baseline method, which corresponds to our previous approach to building Google's production systems, followed by the proposed method currently used to build training sets and train acoustic models for our production recognizer.

2.1. Baseline method

The data selection procedure of the baseline method is represented in Fig. 1. In this approach we directly rely on the transcripts generated by our production system that operates under runtime constraints dictated by real-time user experience considerations. Our average utterance duration is 3.3 seconds, and we start by randomly selecting 50M utterances from our log stream. We filter this set by removing utterances with transcripts shorter than a threshold which is usually 10 characters, but is adjusted on a per-language basis to account for e.g., Asian language character sets. This filtering is based on our empirical observation that confidence scores are less reliable in these cases. After that, we keep the 2M top utterances by confidence.

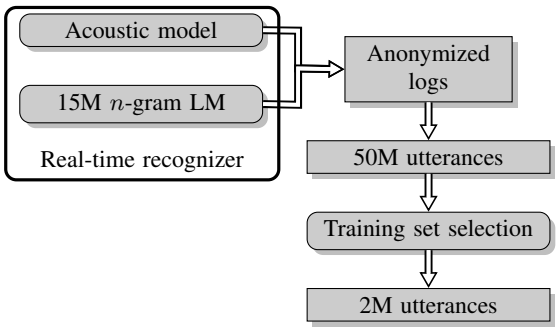

Figure 1: Main components of the baseline method

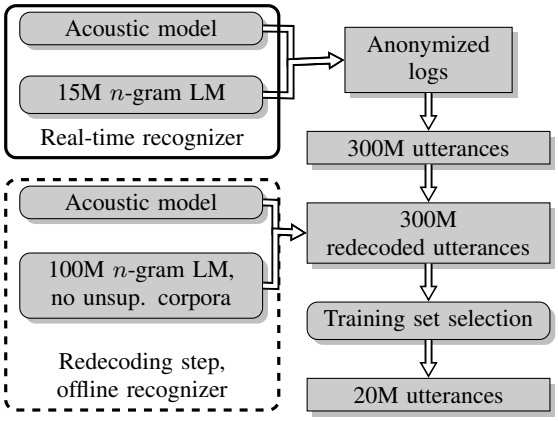

Figure 2: Main components of the proposed method

We then build a DNN acoustic model of medium size - around 6K to 8K context-dependent states, depending on the language (see section 3 for a detailed description of the model). Our DNN models are composed of up to 8 layers of 2560 nodes each. They are trained with our distributed parallelized DNN training toolkit [9].

Our language models (LM) are Katz-smoothed 5-gram models with 1M word vocabulary and 15M n-grams built using the infrastructure described in [10]. Our LMs are trained on corpora composed of different sources, such as books, web documents, search queries and automatically transcribed speech logs.

2.2. Proposed method

Fig. 2 shows the components of the proposed method. We start by extracting 300M utterances, or more than 30 years of speech data, randomly selected from the logs, and redecode them with an offline recognizer. We improve transcription quality in this recognizer in two ways. First we relax many of the decoder parameters such as beam width and number of active arcs during search to reduce expected word error rate at the cost of increasing runtime. Secondly we use a larger language model. We replace the usual production 15M n-gram model by a 100M n-gram language model. In our experiments we have explored even larger sized n-gram language models (up to 1B n-grams) but haven't observed any significant improvements over the 100M n-gram model.

This redecoding pass results in a new corpus of transcripts

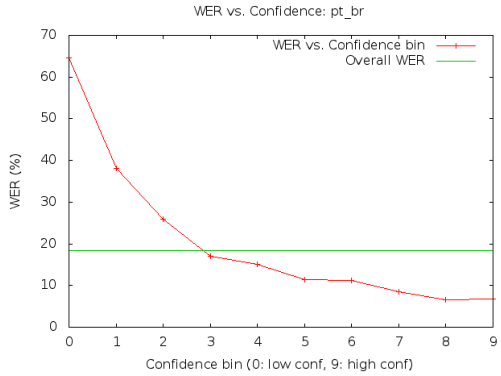

Figure 3: WER vs utterance level confidence bin for Brazilian Portuguese.

with a substantial reduction in expected word error rate. In Table 1 we show the WER reductions of the redecoding system over the baseline production system, when evaluated on human-transcribed test sets. The redecoding system on average performs approximately 9% better relative to the production system. The additional computational overhead incurred to produce higher-quality transcripts for our training corpora is mitigated by our use of Google's massive compute engine infrastructure. By using 8,000 machines we reprocess three decades of speech in approximately 24 hours.

Language	WER′	WER″	Rel. gain	xRT″/xRT′
Russian	27.5	26.0	5.5%	4.6
French	16.2	14.4	11.6%	2.1
Italian	13.6	12.4	8.8%	3.8
Brazilian Portuguese	24.3	23.7	10.5%	4.1

Table 1: Performance comparison of the production (WER′, xRT′) and redecoding systems (WER″, xRT″)

Our next step is to select 20M utterances within this 300M set, a ten-fold increase in training set size over the baseline approach. In this selection we try to guarantee correct transcriptions while preserving the acoustic and linguistic diversity of the data. To guarantee good quality in the transcriptions we rely on utterance level confidence metrics derived from lattice posterior word probabilities [11]. Experimentally we have found that utterance confidence measures correlate extremely well with word error rate. In Figure 3 we show word error rate distributed by confidence score quantiles on a 25 hour Brazilian Portuguese voice search test set. Typically utterances in the 90th percentage confidence bin exhibit word error rates below 10%. In our studies we have found this to be close to the performance of human transcribers.

However, selecting training corpora by confidence score alone may bias the data selection towards frequent and easy to recognize queries. We observed that this can lead to a disproportionately large fraction of the training set having the same transcription. Very popular queries such as 'facebook', 'microsoft', 'google', or 'youtube' can dominate the resulting training set resulting in a negative feedback loop.

To alleviate this problem we have experimented with a number of techniques. For example in [12] an additional constraint is introduced to minimize the cross entropy between the

CD state distribution of the selected training corpora and a manually transcribed testing set. In the present work we opt for a simpler approach that enforces data diversity and is much more straightforward to implement. We limit the number of utterances in the corpus with the same transcription to 20, an empirically determined threshold, an approach we refer to as "transcription flattening". This simple idea enforces a more uniform triphone coverage and a larger vocabulary, while also ensuring that frequent tokens are not over represented.

We also apply additional constraints on the data selection, such as removing utterances with short transcripts as described in Section 2.1. Finally, we retain the top 20M utterances with the highest confidence scores.

3. Training system description

We first derive a context dependent (CD) state inventory using decision tree state clustering as is standard for training Gaussian mixture models [13]. To build this tree we use the entire 20M utterances corpus available for training, since we want to cover as many triphone contexts as possible.

Our acoustic models can contain around 18K CD clustered states, with slight variation by language. While we have observed lower word error rates with even bigger inventories [14], the resulting increase in CPU load is incompatible with real time production systems. Experimentally we find that an inventory of this size is a good compromise.

For each language we train a feed-forward fully-connected neural network [15]. The input layer takes as input 26 consecutive audio frames represented each by a 40-dimensional vector composed of log mel filterbank energies. These 26 audio frames consist of 20 neighboring frames in the past, 5 neighboring frames in the future, and the current frame. We use 8 hidden layers of 2560 nodes each, and an approximately 18K-dimensional softmax output layer. All hidden layer nodes use a rectified linear function [16]. Output layer nodes on the other hand use a softmax non linearity and provide an estimate of the posterior probability of each CD state.

For training we use mini batch asynchronous stochastic gradient descent implemented in a distributed framework [9]. We use a cross entropy criterion as the objective function. The initial learning rate is set to 0.001 and is decreased by a factor of 10 every 5 billion frames.

4. Experimental Results

In what follows, we will first present the results of two controlled experiments assessing the impact of two crucial factors in our final quality gains: improved transcript quality and increased training corpus size. We then report on the total impact of combining all the improvements over the baseline setup as described in Section 2.2.

We first explore the impact of reference transcription quality on the final system. We selected 5M highest-confidence utterances following the procedure described in Section 2.2. We then trained two systems, one using the improved transcripts generated by our new approach, and the other using the original production quality transcripts. Therefore the only difference between these two systems is in the quality of the auto generated transcripts that are aligned against the feature stream to provide training data for the DNN (Table 1 documents the expected transcript quality improvement).

Please note that the absolute quality numbers for a given language differ between the various experiments presented. Since the experiments were not conducted at the same time, this variation is explained by quality improvements made in the time elapsed between experiments being conducted. However, each experiment was performed in a consistent manner, with all aspects of the system being the same except for the improvement being studied.

Table 2 illustrates our results on two different languages: French and Russian. We used a 25 hour test set for each language. The DNNs for each of the systems were trained under similar condition as described in Section 3. We can observe that the use of an improved auto generated reference transcript leads to reductions in the word error rate of the system.

Language	Production transcripts	Improved transcripts	Rel. gain
Russian	27.3	26.6	3%
French	15.4	15.1	2%

Table 2: Effect of the quality of reference transcript on WER

Next we evaluate the impact of the size of the training data set on recognition performance. We selected 3 training sets for three of our production languages. For each language the training sets contained 5, 10, and 20 million utterances, all of which were selected by taking the highest-confidence utterances after redecoding and applying the heuristics described in Section 2.2. We then proceeded to train DNN acoustic models under similar conditions for all data sets and languages.

Table 3 shows our results. Note that a 5M-utterance training run was only available for Russian, as a full exploration of the parameter space for the other languages was not feasible due to time constraints. In general the improvements are small but consistent. Comparing results as the training set size grows, we observe average WER reductions of 0.6% absolute across the three languages. While this might seem small, the WER reductions are statistically significant in the context of our typical 25 hour test sets. In addition, given the scale of Google's voice search traffic, our experience has been that even such small WER reductions translate into significant increases in user satisfaction metrics such as click through rates, lower retries, and lower rejections, among others.

To further compare the quality disparity between the 5M and 20M systems we conducted a human mediated side by side comparison. In this type of test, audio from anonymized recent production logs is reprocessed with the two systems. Those utterances where the transcripts are different are sent to human raters which decide which system produced the better transcript. This type of evaluation pays more attention to the harder to recognize queries and is a good indicator of the underlying quality of the system on real traffic. The side by side test essentially measures whether the difference in performance of two systems is statistically significant.

In this experiment, human raters assign to each transcript i a coarse correctness metric $s_i \in [0, 1]$: 0 means a nonsense transcript, while 1 corresponds to exact recognition result. If n is the total number of utterances sent to raters, we compute $h_i = s_i^{(20)} - s_i^{(5)}$ for $i = 1, \ldots, n$, where $s_i^{(5)}$ and $s_i^{(20)}$ denote the correctness scores assigned to an utterance i recognized by 5M-trained and 20M-trained systems respectively. The side by side test is aimed at rejecting the null hypothesis: that the mean of the difference of the correctness scores between the two systems is 0. To test the null hypothesis, sample mean and bias-corrected variance are first computed based on the correctness

scores provided by the raters, and the null hypothesis is rejected if the calculated p-value is below the chosen significance level α.

In our side by side experiment we sent 500 utterances to human raters and used a standard value of $\alpha = 0.05$. Our test reported a p-value of less than 2%, which conclusively demonstrates that the model trained on the 20M set is preferred by the human raters.

Language	5M training set	10M training set	20M training set
Russian	27.0	26.8	26.2%
Brazilian Portuguese	–	18.2	18.0%
French	–	15.4	14.9 %

Table 3: Effect of unsupervised training set size for various languages: WER for systems trained on 5M, 10M and 20M corpora

Finally, Table 4 compares the baseline data selection approach with our proposed method. We show results across four languages. In all cases word error rate reductions of up to 14% relative have been obtained using this fully automated and unsupervised procedure. This process allows us to easily update acoustic models in a matter of days. Since the demographics and device channel characteristics of the users of Google's speech products are constantly changing, it is important to refresh our acoustic models with as little human intervention and as often as possible. An automated freshness pipeline also helps ensure that any algorithmic improvements can be propagated promptly to all language systems. Currently we retrain our acoustic models across more than 50 languages on a monthly schedule using the approach described in this paper.

Language	WER baseline	WER proposed	Relative gain
Russian	27.5	25.1	8.7%
French	16.2	14.6	10.4%
Italian	13.6	12.1	11%
Brazilian Portuguese	24.3	20.9	14%

Table 4: Performance comparison of the baseline approach (2M training set, production transcripts) and the proposed approach (20M training set, redecoded transcripts)

5. Conclusions

In this paper we presented a fully unsupervised approach to acoustic model data selection. The techniques described take advantage of the large amount of traffic of Google's speech recognition products. This allows us to select corpora of around 20 million utterances with close to human transcriber quality in a matter of hours. We also introduce a procedure to improve the transcription quality further by using a slower and more accurate offline speech recognition system. The combination of these two approaches yields significant reductions in word error rate. Human mediated side by side comparisons of our old and the proposed approaches consistently show the value of the new method.

6. References

[1] J. J. Schalkwyk, D. Beeferman, F. Beaufays, B. Byrne, C. Chelba, M. Cohen, M. Kamvar, and B. Strope, "Google search by voice: A case study," *Advances in Speech Recognition: Mobile Environments, Call Centers and Clinics*, pp. 61–90, 2010.

[2] G. E. Hinton, L. Deng, D. Yu, G. E. Dahl, A. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, 2012.

[3] D. Yu, L. Deng, and G. E. Dahl, "Roles of pre-training and fine-tuning in context-dependent DBN-HMMs for real-world speech recognition," in *NIPS*, 2010.

[4] K. Yu, M. J. F. Gales, L. Wang, and P. C. Woodland, "Unsupervised training and directed manual transcription for LVCSR," *Speech Communication*, vol. 52, no. 7-8, pp. 652–663, 2010.

[5] J. Z. Ma and R. M. Schwartz, "Unsupervised versus supervised training of acoustic models," in *INTERSPEECH*, 2008, pp. 2374–2377.

[6] Y. Huang, D. Yu, Y. Gong, and C. Liu, "Semi-supervised GMM and DNN acoustic model training with multi-system combination and confidence re-calibration." in *INTERSPEECH*. ISCA, 2013, pp. 2360–2364.

[7] H.-K. J. Kuo and V. Goel, "Active learning with minimum expected error for spoken language understanding," in *INTERSPEECH*, 2005, pp. 437–440.

[8] L. Lamel, J. Gauvain, and G. Adda, "Lightly supervised and unsupervised acoustic model training," *Computer Speech and Language*, vol. 16, no. 1, pp. 115–229, 2002.

[9] J. Dean, G. Corrado, R. Monga, K. Chen, M. Devin, Q. V. Le, M. Z. Mao, M. Ranzato, A. W. Senior, P. A. Tucker, K. Yang, and A. Y. Ng, "Large scale distributed deep networks," in *NIPS*, 2012, pp. 1232–1240.

[10] T. Brants, A. C. Popat, P. Xu, F. J. Och, and J. Dean, "Large language models in machine translation," in *In EMNLP*, 2007, pp. 858–867.

[11] H. Jiang, "Confidence measures for speech recognition: A survey," *Speech Communication*, vol. 45, no. 4, pp. 455–470, 2005.

[12] O. Siohan, "Training data selection based on context-dependent state matching," in *ICASSP*, 2014.

[13] S. J. Young, J. J. Odell, and P. C. Woodland, "Tree-based state tying for high accuracy acoustic modelling," in *Proceedings of the Workshop on Human Language Technology*, Stroudsburg, PA, USA, 1994, pp. 307–312.

[14] H. Liao, E. McDermott, and A. Senior, "Large scale deep neural network acoustic modeling with semi-supervised training data for youtube video transcription," in *Workshop on Automatic Speech Recognition and Understanding (ASRU)*, 2013.

[15] N. Jaitly, P. Nguyen, A. W. Senior, and V. Vanhoucke, "Application of pretrained deep neural networks to large vocabulary speech recognition," in *INTERSPEECH*, 2012.

[16] M. Zeiler, M. Ranzato, R. Monga, M. Mao, K. Yang, Q. Le, P. Nguyen, A. Senior, V. Vanhoucke, J. Dean, and G. Hinton, "On rectified linear units for speech processing," in *38th International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vancouver, 2013.