HIGH QUALITY AGREEMENT-BASED SEMI-SUPERVISED TRAINING DATA FOR ACOUSTIC MODELING

Félix de Chaumont Quitry, Asa Oines, Pedro Moreno, Eugene Weinstein

Google Inc.

{fcq, asaj, pedro, weinstein}@google.com

ABSTRACT

This paper describes a new technique to automatically obtain large high-quality training speech corpora for acoustic modeling. Traditional approaches select utterances based on confidence thresholds and other heuristics. We propose instead to use an ensemble approach: we transcribe each utterance using several recognizers, and only keep those on which they agree. The recognizers we use are trained on data from different dialects of the same language, and this diversity leads them to make different mistakes in transcribing speech utterances. In this work we show, however, that when they agree, this is an extremely strong signal that the transcript is correct. This allows us to produce automatically transcribed speech corpora that are superior in transcript correctness even to those manually transcribed by humans. Furthermore, we show that using the produced semi-supervised data sets, we can train new acoustic models which outperform those trained solely on previously available data sets.

Index Terms— semi-supervised, agreement-based, ensemble, data selection, acoustic modeling

1. INTRODUCTION

The performance of a machine learning system is only as good as the data it is trained on. Acoustic models for large-vocabulary continuous speech recognition systems, in particular, must be trained on a large corpus of audio recordings (utterances) annotated with ground-truth transcripts. The performance of the trained model depends on the volume of training data, the accuracy of these transcripts, and how well the training data represents real usage conditions [1].

Traditionally, training corpora for acoustic modeling have been generated in one of two ways. To produce supervised training sets, utterances are given to human transcribers who listen to the data and provide manual transcriptions. This approach has the benefit of yielding highly accurate transcripts, but the process is expensive and time consuming. To produce semi-supervised training sets, utterances are transcribed using existing ASR systems, which generate transcripts and their associated confidence scores.

Various learning approaches have been proposed to use this type of data [1, 2, 3, 4, 5, 6]. Most prior work has been focused on selecting the best utterances to be used for training without modifying existing transcriptions. Speech utterances selected by applying confidence score thresholds and/or other heuristics are kept, with the transcript provided by the ASR system taken as the ground-truth for training purposes. Additionally, in previous work, we explored re-transcribing large corpora of speech data with more powerful recognizers that would not have met the real-time performance restrictions for use in our primary systems [7], and showed that this can produce gains beyond simply using the original transcript provided by the real-time system.

These approaches have the advantage of being inexpensive and straightforward to scale to a large volume of data but the accuracy of the transcriptions is inferior to those produced by human raters. Furthermore, confidence-based selection can bias the data set towards utterances that are "easier" for the system to transcribe correctly, which can lead to a degradation in performance of acoustic models trained on sets generated in this manner [4]. As a result, even with such approaches we still found that the quality of systems trained on manually-transcribed speech data easily surpassed that of systems trained on corpora produced with semi-supervised algorithms. This motivated us to continue to explore methods of automatically creating high-quality transcripts for acoustic model training corpora.

In this paper we propose an agreement-based technique leveraging existing acoustic models to programatically produce high quality semi-supervised transcripts. With this technique we are able to efficiently generate large corpora of utterances with accurate transcripts that can be used for acoustic model training. In section 2 we further describe existing data selection techniques for acoustic modeling and their shortcomings. In section 3, we outline variations of our agreement-based selection technique for producing semi-supervised transcripts. In section 4 we demonstrate that acoustic models trained on data selected using our agreement-based technique have superior performance to those trained on data generated using confidence-based selection. Finally, in section 5, we summarize our results and propose future work.

2. BACKGROUND

We assume that we have access to a large pool of untranscribed speech recordings. In order to make use of such recordings usable for acoustic modeling, we need to transcribe them, either manually (supervised setting), or automatically (semi-supervised).

In the semi-supervised setting, transcripts are created by decoding utterances with an ASR system.[1] This could be simply the same production system as that serving user traffic in real-time, or a slower and more accurate system. Because not all transcripts produced are accurate, we filter the corpus of transcribed utterances to select those with transcripts that are more likely to be correct.

The supervised transcription approach scales poorly - to increase the output volume and/or speed of manual transcription efforts, one needs to employ more and more human transcribers. In contrast, semi-supervised approaches are straightforward to scale to produce

[1] Some authors prefer the term *unsupervised*. We think semi-supervised is more accurate, as pre-existing ASR systems are usually trained on some supervised data, and also because the automatically provided transcript is still a form of supervision.

very large speech corpora quickly and cheaply, as long as adequate computational resources are available.

2.1. Confidence-based selection

The traditional approach (e.g., [2, 4, 6]) to semi-supervised data selection is to use a confidence model that estimates, for each redecoded utterance, a confidence score for the automatically produced transcript.

Given a pool of redecoded utterances and their associated confidences, one can set a confidence threshold to determine what data to keep for training. Certain trade-offs are implicit in the choice of a confidence threshold:

- a high confidence threshold may exclude a large amount of good data and bias the resulting training set towards the easier audio examples;
- a low confidence threshold will decrease the overall accuracy of the produced data set, which may hinder the performance of acoustic models trained on this data.

3. AGREEMENT-BASED AUTOMATIC TRANSCRIPTION

We propose an agreement-based approach as an alternative to traditional confidence-based selection.

Given an ensemble of N ASR systems, we use each system to transcribe unlabeled utterances. The transcripts generated by these decodes may be correct or contain errors. In this work, we show that due to the fact that each individual ASR system was trained on diverse data (e.g., different dialects of the same language or even the same language), it is unlikely that, for a given utterance, they generate transcripts with the same mistakes. We do not expect that this assumption holds up universally, since the mistakes made by each recognizer are not completely independent. For example, the presence of a word mispronounced to sound like another word increases the likelihood of a substitution error across all considered systems. The experimental results presented in Section 4 confirm that system agreement in transcribing speech is an extremely strong signal that the generated transcript is correct. Let K be an agreement threshhold, where $1 < K \leq N$. We propose the following simple agreement-based automatic transcription and selection algorithm.

1. For each unlabeled utterance, we recognize it using N different ASR systems
2. If K of the N transcripts produced agree on the newly produced transcript, we add the utterance and newly associated transcript to our training set.
3. Otherwise we discard it.

3.1. Agreements across dialects

Our proposed agreement-based transcription and selection approach relies on the diversity of the recognizers in the ensemble, in that it is unlikely that several recognizers will make the same mistake. This allows us to substantially reduce transcription errors when compared to simply using a single recognizer.

The first type of ensemble we considered was based on dialects of the same language. For example, English recognizers specialized for different regions of the world. The idea behind this choice is that such dialects are close enough that the majority of the language should be understandable by all recognizers, but diverse enough that errors should be different, thus affording a high likelihood that any agreement between them is evidence of a correct transcription.

One potential drawback of this technique is that it introduces a certain bias in the linguistic content observed in the transcripts produced. This is because words only belonging to a single dialect will never be chosen by agreement with the other dialects. More generally, words which have unusual usage in a given dialect have a lower chance of being recognized correctly by the recognizers of the other dialects, thus lowering the overall chance of having dialect-specific utterances in our semi-supervised corpus.

3.2. Agreements across close languages

In case we don't have multiple dialect systems for a given language, we can instead use close languages. We have experimented with transcribing utterances using the ASR systems of two different languages, and selecting those where both transcripts agree. This approach is of course limited to settings where the target languages are reasonably close to each other.

The same drawback as in 3.1 is even more present: only words that exist in all the languages used can appear in the selected transcripts, thus biasing against words that are specific to either language.

Our results in 4.4 show that this bias is surmountable for acoustic modeling, presumably as long as the entire phone inventory is sufficiently covered.

4. EXPERIMENTAL RESULTS

Our experimentation consisted of transcribing large quantities of anonymized audio logs with multiple speech recognition systems, selecting datasets using the utterances where agreement was observed, and training acoustic models on the data sets constructed in this fashion. We also compared training on various combinations of supervised and semi-supervised data selected both via agreement approaches as well as confidence heuristics.

4.1. Data set creation

Using the proposed agreement-based transcription and selection technique we extracted from anonymized audio logs eight training sets across three different languages: English, Arabic, and Malay. An ensemble of US, British, and Australian English dialect systems was used to create the English training sets. An ensemble of Egyptian, Levantine, Maghrebi, and Gulf Arabic dialect systems was used to create the Arabic training sets. Finally, an ensemble of Malay and Indonesian systems was used to create the Malay training set, as described in 3.2. It is important to point out that while the present work is concerned only with the training of acoustic models, the dialect and language systems just mentioned did not only differ from each other in the acoustic models used, but also had disparate language models trained on country/dialect specific text sources, and, in some cases, different lexicons and other individual modifications specific to the target language or dialect.

Table 1 describes the training sets created by this technique.

In our experimentation, we observed that the agreement rates varied depending on the language. For English dialects, 3-out-of-3 agreement was obtained on around 20% of the data processed, while for Arabic languages, the 3-out-of-4 agreement rates were closer to 30%. Finally, the Malay-Indonesian combination had a 2-out-of-2 agreement rate of 14%.

Language	Recognizers	Origin	K	Size
English	Australia, Great Britain, United States	Great Britain	3	9M
		Australia	3	6M
		Uganda, Tanzania	3	365k
		Philippines	3	5M
		South Africa	2	1.3M
Arabic	Gulf, Levantine, Egyptine, Maghrebi	Egypt	3	1.5M
		Levant	3	3M
Malay	Malay, Indonesian	Malaysia	2	768k

Table 1: Training sets created with the agreement-based method

4.2. Evaluating Agreement Rates and Data Set Quality

We sent US English utterances sampled from various semi-supervised data sets, using mixes of both confidence-based and agreement-based techniques, as well as US English utterances from a supervised training set, to have their transcripts marked as correct or incorrect by human raters. The results of these ratings are shown in table 2. From these results, we see that training utterances obtained via 3-out-of-3 agreement between the individual recognizers are much more likely to be transcribed correctly than both semi-supervised data selected using confidence thresholds as well as manually-transcribed supervised utterances.

Additionally, we can see that utterances with low confidence that were created from full agreement also have a higher correctness percentage. This result suggests that agreement-based transcription and selection is an effective way of avoiding the potential bias arising from selecting only high-confidence utterances, which is a concern in conventional approaches for semi-supervised acoustic model training data selection.

Figure 1 shows the confidence histograms of an agreement-based semi-supervised data set. This shows that while the agreement-based technique does favor higher confidence utterances, that is to say those which are presumably easier to recognize, it also includes a substantial number of low-confidence utterances. This is a clear advantage over the confidence-based selection technique, which, by definition, will not include any. These plots show that when it is a concern that selecting a data set by agreement yields a biased sample of the data, it is possible to discard a random sample of the higher-confidence utterances so as to recover the original confidence distribution.

4.3. Experimental Setup

Using the data sets described in 4.2, we proceeded to train a variety of models in the respective languages. For each language, we compared our current best model with one trained using the newly extracted data.

4.3.1. Model architectures

In our work, we evaluated two different types of acoustic modeling architectures, both based on long short-term memory (LSTM) recurrent neural networks [8, 9]. This is due to the fact that we don't usually roll out an algorithmic improvement to all languages at the same time, and hence various languages were in various stages of algorithmic upgrades at the time of the experiments. It is important to point out, however, that within a specific language the experiments we describe below are consistent - that is, we always use the same architecture for any given language or dialect.

Here are the two architectures that we used for our experiments:

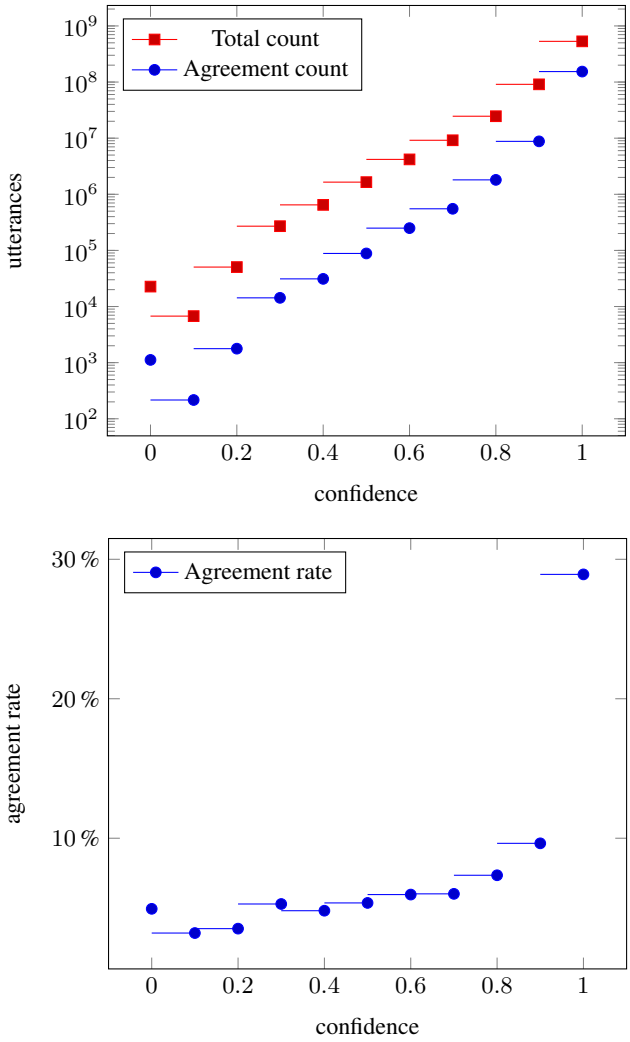

Fig. 1: Utterance counts and agreement rate by confidence

CLDNN A variation of the architecture described in [10], which consists of:

- 1 Convolution layer taking 40-dimensional filterbanks vectors as input, with projection to 256 units;
- 3 LSTM layers with 832 units each, with recurrent projections to 512 units;
- 2 Feed-forward layers with ReLU activations, respec-

Data set type	Filtering used		Utterances	
	Agreement-based	Confidence-based	Sample size	Correct
Supervised	N/A	N/A	1496	**89%**
Semi-supervised	None	$conf \geq 0.9$	4888	**88%**
	3-out-of-3	None	5000	**97%**
		$conf \geq 0.9$	4674	98%
		$conf < 0.9$	326	94%
	2-out-of-3	$0.5 \leq conf < 0.9$	4000	70%
		$conf < 0.5$	1134	54%

Table 2: Utterances sent for validation

tively with 1024 and 512 units;

- 1 Softmax layer classifying into context-dependent (CD) states. The dimension of the layer depends on the CD state inventory size of the target language.

CTC A variation of the architecture described in [11], which consists of:

- 5 LSTM layers with 600 units each, no projection;
- 1 CTC layer classifying into CD state sequences. The dimension of the layer depends on the CD state inventory size of the target language.

4.3.2. Data set combinations

Various languages initially had various baseline sets of data. The three situations we encountered were:

1. one large supervised data set (usually 3M utterances);
2. one small supervised data set (around 100k utterances), and one large confidence-based semi-supervised data set (at least 1M);
3. one large confidence-based semi-supervised data set only (at least 1M);

For case 1., we experimented with supplementing the supervised set with our agreement-based semi-supervised set.

For cases 2. and 3., we tried replacing the confidence-based set with the agreement-based set if the sizes were comparable, or simply supplementing with our agreement-based set if significantly smaller than its confidence-based counterpart (e.g., twice as small).

In the following we will use the following abbreviation for each type of data set:

sup. for supervised data sets;

conf. for confidence-based semi-supervised data sets;

agree. for agreement-based semi-supervised data sets.

4.3.3. Training methods

Each model was trained using a sequence of two training methods.

The first method depends on the model architecture: for CLDNN models we used a cross-entropy (CE) criterion [8], for CTC models we used the CTC loss [12].

The second method was the same for both architectures: we used discriminative sequence training using sMBR [9, 11].

All training used multi-style training (MTR) with the same reverberation and noise configurations, both for noise-robustness and regularization.

4.4. Word Error Rate Results

For each confidence-based semi-supervised data set, we compared the quality of the overall system using an acoustic model trained with and without the new set. The exact data sets and their sizes used for each experiment are detailed in 3.

Quality changes were measured on two types of test sets:

VS a set of voice search utterances, in the given language;

IME a set of dictation utterances, in the given language.

Those sets typically include between 2k and 10k utterances, depending on the language.

Table 3 presents our results in terms of word error rate for the various languages, on both types of test sets (no IME test set was available for two of the languages).

Subtable 3a shows how replacing confidence-based semi-supervised sets with agreement-based sets can yield substantial quality gains. In the case of Philippine English we didn't see gains as large as in the other cases. One reason for this could be that this dialect of English is not as distinct from US English as in the cases of African varieties (South Africa, Nigeria), and as a result it is easier to transcribe utterances from the Philippines correctly with even a single US English recognizer in order to produce a high-confidence semi-supervised training data set.

Subtable 3b shows how supplementing confidence-based semi-supervised set with agreement-based semi-supervised sets can yield quality gains, even though the size of the agreement-based set may be much smaller than its counterpart.

Finally, subtable 3c shows how supplementing supervised sets with large confidence-based semi-supervised sets can yield quality improvements.

5. CONCLUSION

We have described a new technique for automatically transcribing and selecting acoustic model training data corpora from a large set of speech audio recordings. This agreement-based selection technique assumes the existence of diverse recognizers that are able to transcribe speech in the target language.

Using this technique, we have produced semi-supervised data sets for several languages, and in turn have used those data sets to train new acoustic models. We have demonstrated both through manual evaluation as well as through quality results of the final systems, that the data sets are of extremely high quality. Additionally, by augmenting supervised data sets and replacing semi-supervised ones selected with conventional heuristics we have been able to train better quality acoustic models than was possible with the previously-available data sets.

Language, Dialect	Model	Data sets	Sizes	VS		IME	
				WER	rel. gains	WER	rel. gains
Arabic, Levantine	CLDNN	conf.	3M	19.9		26.3	
		agree.	3M	17.2	-13.6%	23.5	-10.6%
English, South Africa	CLDNN	conf.	1.6M	21.5		22.9	
		agree.	1.3M	18.6	-13.5%	19.5	-14.8%
English, Philippines	CTC	sup. + conf.	160k + 3M	18.2			
		sup. + agree.	160k + 5M	18.1	-0.5%		N/A

(a) Replacing confidence-based sets with agreement-based sets of similar size

Language, Dialect	Model	Data sets	Sizes	VS		IME	
				WER	rel. gains	WER	rel. gains
English, Nigeria	CLDNN	sup. + conf.	80k + 1.3M	31		27.5	
		sup. + conf. + agree.	80k + 1.3M + 360k	29.2	-5.8%	26.2	-4.7%
Arabic, Egypt	CLDNN	conf.	3M	26		29.3	
		conf. + agree.	3M + 1.5M	25	-3.8%	28.6	-2.4%
Malay, Malaysia	CTC	conf.	1.5M	29.3		36	
		conf. + agree.	1.5M + 760k	27.4	-6.5%	33.4	-7.2%

(b) Supplementing confidence-based sets with smaller agreement-based sets

Language, Dialect	Model	Data sets	Sizes	VS		IME	
				WER	rel. gains	WER	rel. gains
English, Australia	CTC	sup.	3M	13.6			
		sup. + agree.	3M + 6M	12.5	-8.1%		N/A
English, Great Britain	CTC	sup.	3M	13.6		12.5	
		sup. + agree.	3M + 9M	12.9	-5.1%	12.3	-1.6%

(c) Supplementing supervised sets with large agreement-based sets

Table 3: Experimental word error rates

Future work includes finding ways to adapt this technique to the situation where the language of interest has no existing dialect variants (or close enough languages) that can be used to readily form an ensemble of recognizers.

6. REFERENCES

[1] Jeff Z. Ma and Richard M. Schwartz, "Unsupervised versus supervised training of acoustic models," in *Interspeech*, 2008, pp. 2374–2377.

[2] Ngoc Thang Vu, Franziska Kraus, and Tanja Schultz, "Multilingual a-stabil: A new confidence score for multilingual unsupervised training," in *Spoken Language Technology Workshop*, 2010, pp. 183–188.

[3] Kai Yu, Mark J. F. Gales, Lan Wang, and Philip C. Woodland, "Unsupervised training and directed manual transcription for lvcsr," *Speech Communication*, vol. 52, no. 7-8, pp. 652–663, 2010.

[4] Yan Huang, Dong Yu, Yifan Gong, and Chaojun Liu, "Semi-supervised gmm and dnn acoustic model training with multi-system combination and confidence re-calibration.," in *Interspeech*. 2013, pp. 2360–2364, ISCA.

[5] Hong-Kwang Jeff Kuo and Vaibhava Goel, "Active learning with minimum expected error for spoken language understanding," in *Interspeech*, 2005, pp. 437–440.

[6] Pengyuan Zhang, Yulan Liu, and Thomas Hain, "Semi-supervised dnn training in meeting recognition," in *Spoken Language Technology Workshop*, 2014, pp. 141–146.

[7] Olga Kapralova, John Alex, Eugene Weinstein, Pedro Moreno, and Olivier Siohan, "A big data approach to acoustic model training corpus selection," in *Interspeech*, 2014.

[8] Haşim Sak, Andrew Senior, and Françoise Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Interspeech*, 2014.

[9] Haşim Sak, Oriol Vinyals, Georg Heigold, Andrew Senior, Erik McDermott, Rajat Monga, and Mark Mao, "Sequence discriminative distributed training of long short-term memory recurrent neural networks," in *Interspeech*, 2014.

[10] T. Sainath, O. Vinyals, A. Senior, and H. Sak, "Convolutional, long short-term memory, fully connected deep neural networks," in *ICASSP*, 2015.

[11] Andrew Senior, Hasim Sak, Félix de Chaumont Quitry, Tara N. Sainath, and Kanishka Rao, "Acoustic modelling with cd-ctc-smbr lstm rnns," in *ASRU*, 2015.

[12] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006.