Semi-supervised acoustic model retraining for medical ASR

Greg P. Finley[1], Erik Edwards[1], Wael Salloum, Amanda Robinson[1], Najmeh Sadoughi[1], Nico Axtmann[3], Maxim Korenevsky[1], Michael Brenndoerfer[2], Mark Miller[1], and David Suendermann-Oeft[1]

[1] EMR.AI Inc., San Francisco, CA, USA
[2] University of California Berkeley, CA, USA
[3] DHBW, Karlsruhe, Germany
greg.finley@emr.ai

Abstract. Training models for speech recognition usually requires accurate word-level transcription of available speech data. For the domain of medical dictations, it is common to have "semi-literal" transcripts available: large numbers of speech files along with their associated formatted episode report, whose content only partially overlaps with the spoken content of the dictation. We present a semi-supervised method for generating acoustic training data by decoding dictations with an existing recognizer, confirming which sections are correct by using the associated report, and repurposing these audio sections for training a new acoustic model. The effectiveness of this method is demonstrated in two applications: first, to adapt a model to new speakers, resulting in a 19.7% reduction in relative word errors for these speakers; and second, to supplement an already diverse and robust acoustic model with a large quantity of additional data (from already known voices), leading to a 5.0% relative error reduction on a large test set of over one thousand speakers.

Keywords: Medical speech recognition, ASR, medical dictation, acoustic modeling

1 Introduction

Training automatic speech recognition (ASR) systems requires transcribed speech corpora to build acoustic models (AMs) and language models (LMs). Traditionally, such transcriptions are created by human labor, which imposes limitations on how large such corpora can be, how many speakers they can cover, how quickly they can be created, and how consistently transcriptions are following required guidelines. To overcome these limitations, techniques have been proposed to create transcriptions automatically, substantially increasing the size of the training corpus with relatively little effort. For example, Suendermann *et al.* perform speech recognition on millions of utterances collected in industrial spoken dialog systems and determine, based upon the recognizer's confidence score, which of the hypotheses can be accepted without further review and which ones

should undergo human quality assurance [8]. Such fully automatic techniques suffer from the disadvantage that they rely on pre-existing speech recognition models and settings and have no way to acquire new vocabulary or adapt to new domains. Thus, they suffer if there is a significant mismatch between training and adaptation language.

The medical transcription domain is a special case, however, in that speech recordings of clinical dictations are almost always subject to transcription into a formatted outpatient report which contains a well-formatted and corrected version of the dictated matter. Note that the process of correcting and modifying a literal transcript into a report is an extensive one and often involves changes that make it impossible to use reports directly as ASR training data: intuiting punctuation, list numbering, etc. when formatting is not explicitly spoken; executing requests by the speaker ("scratch that,", e.g.); or even inserting material from elsewhere in the patient's medical history.

Strategies for using this very rich set of data for the purpose of model enhancement, and to overcome its lack of word-level correspondence between spoken and written content, have been discussed in the literature for about two decades. Early research showed that this type of data can indeed be used to adapt a speaker-independent model to new speakers [9, 5]; the basic approach is to use an ASR engine to decode new audio with matching reports, then use the results that can be verified correct as new training data. However, these studies use very small test sets and speech recognition technology which is widely considered dated. Consequently, the baseline performance is very poor by modern standards, and reported improvements often do not meet statistical significance thresholds. To increase the amount of usable data beyond only the correct outputs of the recognizer, researchers have also explored using LMs for decoding built specifically to the report [5] or have explored the use of phonetic [7] and semantic [2, 6] features to correct ASR errors using the report as reference. However, the latter studies either did not test how accuracy of a speech recognizer is impacted when adding the new data, or limited the study to LM adaptation.

Outside of the medical domain, this type of semi-supervised approach has more recently been applied to parliamentary transcription, which is a similar case in that large amounts of semi-verbatim transcription data are available [3, 4]. To our knowledge, however, no validation of these methods exists for building AMs for a modern, production-scale medical ASR system. In this paper we present such a validation for two applications: adapting a model to previously unseen speakers, and enhancing an already large model with additional data from known speakers.

2 Method

We applied semi-supervised methods to enhance the training corpus for AMs in two different experiments. Experiment 1 represents a case of speaker adaptation, using semi-supervised data for speakers unknown to the original acoustic model. In Experiment 2, on the other hand, we test whether a model can be augmented

by adding a large quantity of additional data from many known speakers. The general procedure for both experiments is the same; they differ only in the data sources used. Except where otherwise specified, the methodological details given below are identical for both experiments.

For each experiment, we built two AMs and compared their performance in word error rate (WER) on a test set. AM1 was a "traditional" model, trained from fully manually transcribed dictations; AM2 contained all the data of AM1 plus a large set of "virtual transcriptions," generated by 1) ASR decoding of a large set of untranscribed data, then 2) identifying correct hypotheses by comparing with matching reports. The entire training and testing process, including all data and models, is described in detail in 2.2 and visualized in Figure 1.

2.1 Data

The primary source of training data consists of manually transcribed dictations, as do all test sets for results reported in this paper. For Experiment 1, no speakers from Test are represented in Train; for Experiment 2, all speakers in Test have exactly one or two dictations in Train. (Recall that Experiment 1 tests the adaptation of a model to new speakers, and Experiment 2 tests the bolstering of an already comprehensive model with more data.)

In addition, we have access to a large number of audio dictations with corresponding reports but no transcripts. This corpus constitutes the "Untranscribed" set for each experiment. See Table 1 for size statistics of all corpora.

In general, corpora used for Experiment 2 are much larger than for Experiment 1. The data also come from different providers, with different speakers, recording conditions, and report styles. Despite the methodological similarity between the two experiments, they should be considered entirely separate cases.

Table 1. Summary of all dictations. Manual transcriptions are available for Train and Test, and reports for Untranscribed. AM1 was trained on Train and AM2 on Train+Aug.

Data set	# speakers	# utterances	# hours
Experiment 1			
Train	245	6,857	305.0
Test	26	32	3.5
Untranscribed	458	12,207	652.8
Augmentation	457	211,909	259.5
Train+Aug.	702	218,766	564.5
Experiment 2			
Train	2,384	9,214	396.1
Test	1,033	1,033	28.9
Untranscribed	1,241	93,581	6646.5
Augmentation	1,228	2,269,801	2617.1
Train+Aug.	2,384	2,279,015	3013.2

Although manual transcriptions are generally considered to be the most accurate source of data for ASR training, medical speech is notoriously difficult due to a number of factors including specialized vocabulary, high rate of speech, etc. [1]. The medical transcriptionists who created Train and Test did so with the aid of matching reports, which themselves were generated through multiple rounds of transcription and quality assurance by other trained transcriptionists.

Additionally, to estimate human WER when unaided by reports, we obtained separately three rounds of transcription on a set of 334 dictations: two rounds using reports as a reference, as is our normal procedure, and one "blind" round. These dictations did not overlap with any other data set. Note also that these reports were taken from the same provider as the data from Experiment 2, so any human WER results should only be considered relevant to Experiment 2.

2.2 Generating additional AM training data

The entire Untranscribed set was decoded using our best prior acoustic model and a specially designed language model (AM1 and LM1, described below). Sequences of correctly recognized words in the hypotheses were identified by aligning hypotheses with reports using a dynamic programming algorithm. Any sequence consisting of five or more consecutive words matching perfectly between hypothesis and transcript was excised, alongside its matching audio range, and considered a training utterance in a large set of supplementary, semi-supervised training data, which we call the Augmentation set (see Table 1). We decided upon a five-word window based on an informal assessment of the excised clips; shorter windows exhibited more slight errors in word boundary detection, which we suspected would propagate in re-training.

Our approach for generating training data is conservative in that we only allow perfect matches of substantial length between hypothesis and report. This ensures that virtual transcriptions are as accurate as possible. Although we piloted some strategies for correcting hypotheses using reports, we have found that, for the quanitities of data that we are considering, the perfect matches already provide very large training corpora by themselves.

2.3 Acoustic and language modeling

Our speech recognizer is based on a state-of-the-art stack with 40-dimensional MFCCs, deltas and delta-deltas, fMMLR, ivectors, SAT, GMM-HMM pre-training, and a DNN acoustic model. Two n-gram LMs were used: a trigram model (LM1) for decoding the large Untranscribed set, and a 4-gram model (LM2) for the experimental results comparing AM1 and AM2. (LM1 is faster to decode with, whereas LM2 is more accurate, so LM1 was chosen for the massive Untranscribed set and LM2 to achieve the best possible results on Test.) To generate LM1, language models are first built for 1) the Train set and 2) the Train + Untranscribed sets; these two are then interpolated, with coefficients tuned to minimize perplexity on a held-out set, to yield the final model. The procedure for LM2 was the same, except that all n-gram counts of Untranscribed were

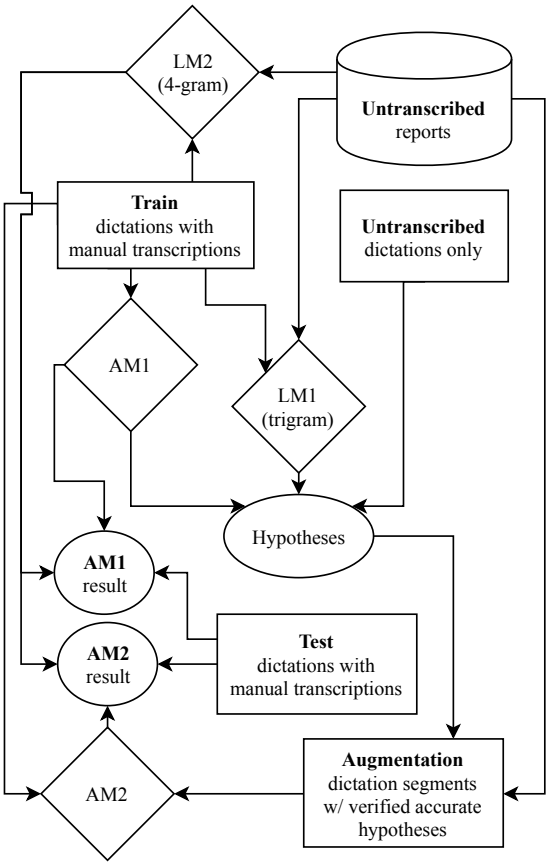

Fig. 1. Experimental training and decoding procedure. Rectangles represent audio, possibly transcribed; cylinders, reports; diamonds, models; ellipses, decoding results.

decremented by one, effectively removing singletons and significantly accelerating decoding for an otherwise slow 4-gram model with minimal effect on WER.

At no point did we use Augmentation data to train LMs. We suspected that doing so would bias the recognizer towards easy speech and very short utterances. (Note also that some version of the linguistic information from the Augmentation set is already present in the LM, which contains Untranscribed.) This bias is not a concern for AM training, where the currency of recognition is at the phonetic level, and transitional probabilities between words are less important.

3 Results: Experiment 1

For Experiment 1, we compared WER on our test set between the baseline acoustic model (AM1) and the large expanded acoustic model (AM2). AM2 decreases the WER from AM1 by 19.7% relative, from 23.1% WER (5,377 edits

out of 23,257 words) to 18.6% (4,317 edits), a statistically significant difference as determined by a test of equal proportions ($\chi^2 = 146.2$, $p < .001$). Out of 26 speakers, 22 exhibit a decline in WER—up to a 52.6% relative reduction in the most extreme case (from 105 errors down to 46 errors, out of 512 words). Of the 4 that see an increase, the highest is an 11.1% relative increase (72 errors up to 80, out of 436 words). The distribution across speakers of relative WER change is visualized in Figure 2.

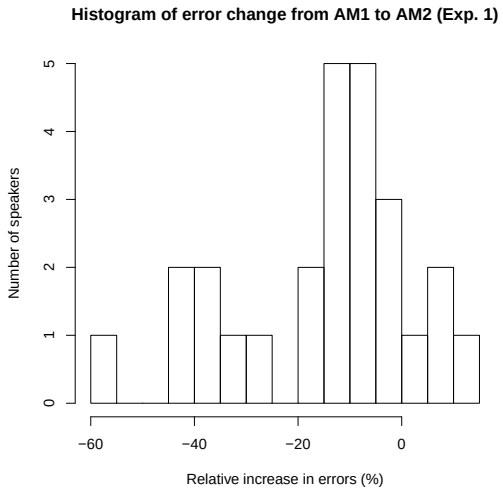

Fig. 2. Relative WER change by speaker, AM1 to AM2 (Experiment 2).

4 Results: Experiment 2

For Experiment 2, we also measured differences in WER between two acoustic models. Additionally, as the test set contains a much larger number of speakers compared to Experiment 1, we dive deeper into the by-speaker results. Note again that *all* models and corpora in this Experiment are different than those used in Experiment 1; mentions of 'AM1'/'AM2' in this section refer now to the Experiment 2 versions of these.

4.1 Decoding accuracy

Decoding with AM2 decreases the WER from AM1 by 5.0% relative, from 22.0% WER (52,961 edits out of 240,382 words) to 20.9% (50,332 edits). Though this effect is smaller than that demonstrated in Experiment 1, the difference is still statistically significant ($\chi^2 = 85.2$, $p < .001$).

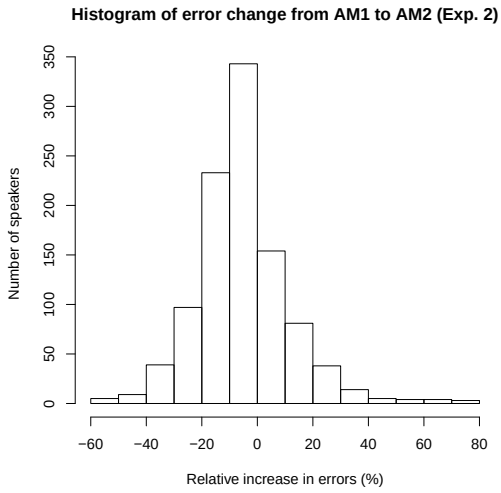

Fig. 3. Relative WER change by speaker, AM1 to AM2 (Experiment 2).

The decrease in error rate is far from uniform across all speakers, however: relative WER over each speaker decreases by as much as 56% and *increased* by as much as 75%. WER increases for 303 out of 1033 speakers. See Figure 3 for the distribution of relative change for individual speakers.

4.2 Effect of amount of data added

The extreme range of variation between speakers, and the fact that many speakers actually see a deterioration in performance, is a surprising finding that invites an explanation. Towards this end, a natural question is whether there is any relationship between the observed changes in WER and the amount of audio data added from the Augmentation set. Across all speakers, there is a correlation between relative change in WER and minutes of audio added, albeit a weak one (Kendall's $\tau = -.046$, $p = .026$; correlation is measured over ranks because time added is non-parametric, with a long right tail). This correlation measurement is only possible given the huge number of speakers in the Experiment 2 test set; no significant similar effect could be observed for Experiment 1.

The relationship between time added and WER is visualized in Figure 4. For this plot, speakers are grouped into bins according to the amount of audio data added, with each bin accounting for a 10-minute range (inclusive on the low end only). The plot shows WER for AM1 (narrow end of the trapezoid) and AM2 (wide end) for each bin—thus, the trapezoid "points" in the direction of the change—calculated over all utterances in that bin. We performed a test of equal proportions for each bin, applying Bonferroni correction for multiple comparisons; those five bins with $p < .05$ are starred in the plot. (Note that the

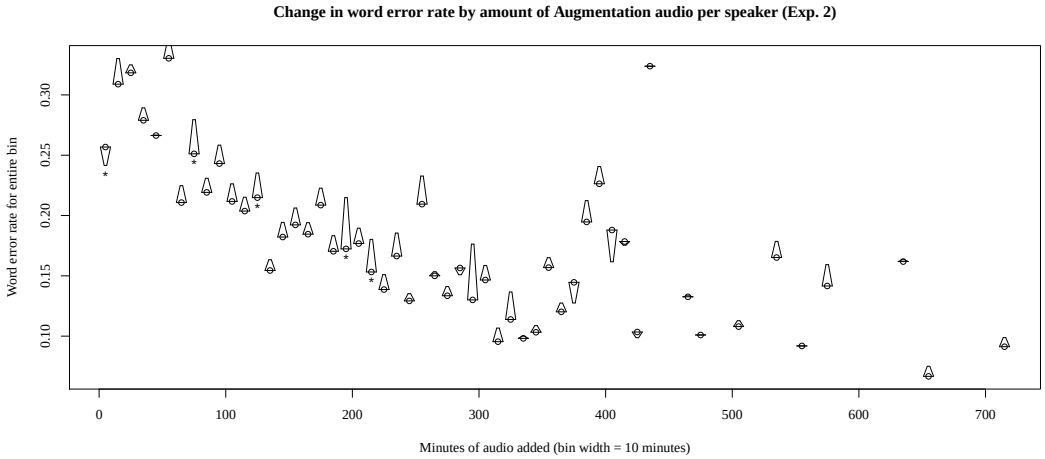

Fig. 4. Binned speaker WERs by amount of audio for each speaker in Augmentation data (Experiment 2). AM1 WER is marked by the narrow end of the bar, AM2 WER by the wide end with a circle. Asterisks underneath bars denote statistical significance of WER change from AM1 to AM2 ($\alpha = .05$, Bonferroni correction).

degree of change in a bin is not necessarily tied to statistical significance, as bins do not all contain the same number of speakers or spoken words.)

These individual bins are rather small, so most do not show statistically significant changes; all those that do are for speakers with fewer than 220 minutes of speech added. Most interesting, however, is that the only bin to show a significant *increase* in WER using AM2 is the 0- to 10-minute bin. This increase is driven mostly by the 30 speakers (out of the bin's 44 total) who had *no* additional data added and saw an increase in WER of over 2% absolute, 8.5% relative ($\chi^2 = 16.2$, $p < .001$). These 30 speakers stand in stark contrast to the dataset as a whole, which shows a 1.1% decrease in absolute WER.

4.3 Human word error rate

Table 2. Human WER between different sets of transcriptions. The "Assisted" conditions were done by professional transcriptionists using matched final reports as a reference, and "Unassisted" by transcriptionists without access to the reports.

Comparison	WER
Assisted1–Assisted2	9.3%
Assisted1–Unassisted	18.0%
Assisted2–Unassisted	20.1%

Given the nature of the data used in this work (recordings with at times extreme noise, non-native accent, audio compression artifacts, hesitations, etc.), and inspired by an earlier publication along these lines [1], we decided to study inter-rater consistency of the dataset by measuring the human error rate. Since our standard transcription procedure (Assisted condition) provides transcriptionists with the existing outpatient report of the dictation (which itself had undergone at least two tiers of transcription), we decided to conduct two types of human error rate experiments: (a) compare two transcriptions of the same audio files created in the Assisted condition and (b) compare transcriptions created in the Assisted condition with those in the Unassisted condition. We expected (a) to exhibit a lower WER than (b) due to the existence of shared material.

The inter-transcriber WERs are given in Table 2. In the Unassisted condition, transcribers differ from the Assisted conditions by 18.0% to 20.1%. From these results, it appears that WER on our data by a single transcriber without pre-generated reference material would approach 20%. Even when such material is available, however, there are notable disagreements or errors in transcription (9.3%), further emphasizing the difficulty of the speech in these dictations. Recall again that we commissioned these transcriptions only for the data used in Experiment 2; human WER for Experiment 1 may not be this high.

5 Discussion

Our proposed method of providing guaranteed accurate data for AM retraining leads to models with lower average decoding error rates. For the purposes of adapting a model to previously unseen speakers, there was a major reduction in WER, eliminating nearly a fifth of all errors. When bolstering an already large model, the gains are somewhat more modest—especially so when considering that AM2 in Experiment 2 was trained on 7.6 times the amount of audio data as AM1. Our human WER measurements do suggest, however, that these dictations are especially difficult, and that we are already approaching human accuracy, so it may simply be the case that performance of the acoustic models has been "saturated" by this point.

The more mixed results in Experiment 2, as well as the large and diverse test set used, invite some speculation as to how speakers may be affected differently *vis-à-vis* their WER by the data augmentation step. Despite the average drop in WER with AM2, performance did deteriorate in some instances. This was most evident for speakers for whom no data was added to the model. We suspect the cause is that the representation of these speakers in AM2 was diluted compared to their representation in the much smaller AM1. As a concrete recommendation, we would not suggest using an augmented acoustic model for speakers who had no data added, assuming they were already represented in the base model.

Other than in this specific case, however, it was difficult to demonstrate any strong relationship between the amount of data added for a speaker and the degree of recognition improvement. One explanation may be the presence of a confounding effect: speakers with higher AM1 WERs will naturally have less

data in the Augmentation set. Because accurate recognition on Untranscribed is a prerequisite for finding utterances to add to Augmentation, speakers for whom the model already does well tend to have the most added data. Indeed, there is a moderately strong correlation (Kendall's $\tau = -.20$, $p < .001$) between AM1 WER and amount of data added per speaker; note that this correlation is visually unmistakable in the general downward trend on the left side of Figure 4. Thus, speakers with the most added data tend to be those who already showed low WER before augmentation. These same speakers would have had less "room for improvement" from changes to the AM: indeed, those speakers with higher AM1 WER tended to have larger relative improvements than those with lower AM1 WER ($\tau = -.065$, $p = .002$). Taken together, the effects of prior AM1 WER on WER change and on amount of data added may be obscuring some of the positive effects of having more added data.

Further gains in performance may be possible via strategies described in the literature for using reports to correct ASR errors on the Untranscribed set, allowing speech previously missed by the recognizer to be used for training. While our methods are sufficient to produce a very large training set, it is likely that adding more difficult speech to training would improve recognition further. This would effectively be an automated active learning approach, using alignment with reports as a semi-supervised step. We also did not attempt to bolster LMs in the same way we did for AMs; however, fully corrected machine transcripts would make this possible to test also.

6 Conclusion

We presented and evaluated a semi-supervised method for augmenting a speaker-independent AM using large numbers of dictations with matching final reports. Our bolstered AMs achieve a significant reduction in error rates, inching closer to human error rates. The methods detailed here are especially effective as a means of adapting an AM to new speakers.

By measuring performance on a large test set of over 1,000 speakers, we were able to note patterns in the procedure's effects. The amount of data added seems not to matter much, except that those speakers without any added acoustic data saw on average an increase in WER. This leads naturally to the conclusion that, whenever practical, different AMs should be used for different speakers depending on whether or not data from the target speaker was added in the augmentation stage.

Future work will include expanding the approach to language modeling and applying more sophisticated techniques to select optimal models, e.g. using speaker clustering. We will also look deeper into the influence of the human error rate on ASR performance in both training and testing cycles and possible techniques to enhance inter-rater reliability for this difficult domain.

References

1. Edwards, E., Salloum, W., Finley, G., Fone, J., Cardiff, G., Miller, M., Suendermann-Oeft, D.: Medical Speech Recognition: Reaching Parity with Humans. In: Proc. of SPECOM. Hatfield, UK (2017)
2. Jancsary, J., Klein, A., Matiasek, J., Trost, H.: Semantics-based automatic literal reconstruction of dictations. Semantic Representation of Spoken Language pp. 67–74 (2007)
3. Kawahara, T.: Transcription system using automatic speech recognition for the japanese parliament (diet). In: IAAI (2012)
4. Kleynhans, N., De Wet, F.: Aligning audio samples from the south african parliament with hansard transcriptions (2014)
5. Pakhomov, S., Schonwetter, M., Bachenko, J.: Generating training data for medical dictations. In: Proceedings of NAACL-HLT. pp. 1–8 (2001)
6. Petrik, S., Drexel, C., Fessler, L., Jancsary, J., Klein, A., Kubin, G., Matiasek, J., Pernkopf, F., Trost, H.: Semantic and phonetic automatic reconstruction of medical dictations. Computer Speech & Language **25**(2), 363–385 (2011)
7. Petrik, S., Kubin, G.: Reconstructing medical dictations from automatically recognized and non-literal transcripts with phonetic similarity matching. In: ICASSP. vol. 4, pp. IV–1125. IEEE (2007)
8. Suendermann, D., Liscombe, J., Pieraccini, R.: How to Drink from a Fire Hose: One Person Can Annoscribe 693 Thousand Utterances in One Month. In: Proc. of SIGdial. Tokyo, Japan (2010)
9. Wightman, C.W., Harder, T.A.: Semi-supervised adaptation of acoustic models for large-volume dictation. In: Proceedings of Eurospeech. pp. 1371–1374 (1999)