An Efficient Repair Procedure For Quick Transcriptions

Anand Venkataraman, Andreas Stolcke, Wen Wang,
Dimitra Vergyri, Venkata Ramana Rao Gadde, Jing Zheng

Speech Technology and Research Laboratory
SRI International
{anand,stolcke,wwang,dverg,rao,zj}@speech.sri.com

Abstract

We describe an efficient procedure for automatic repair of quickly transcribed (QT) speech. QT speech, typically closed captioned data from television broadcasts, usually has a significant number of deletions and misspellings, and has a characteristic absence of disfluencies such as filled pauses (for example, um, uh). Errors of these kinds often throw an acoustic model training program out of alignment and make it hard for it to resynchronize. At best the erroneous utterance is discarded and does not benefit the training procedure. At worst, it could misalign and end up sabotaging the training data. The procedure we propose in this paper aims to *cleanse* such quick transcriptions so that they align better with the acoustic evidence and thus provide for better acoustic models for automatic speech recognition (ASR). Results from comparing our transcripts with those from careful transcriptions on the same corpus, and from comparable state-of-the-art methods are also presented and discussed.

1. Introduction

Speech transcription is a time-consuming and expensive process, often involving expert guidance, supervision and multiple iterations. As a consequence, when large amounts of data are required to be distributed for training of acoustic models, it has become common to distribute quick transcriptions. In QT data, the goal is to get as much data out the door as rapidly as possible, without paying too much attention to the quality of the transcript produced. Obviously, this results in more errors, but the implicit expectation is that the large volume of data will more than make up for any errors in the transcription process. The Linguistic Data Consortium (LDC) has, under this assumption, recently distributed several thousand hours of closed captioned broadcast news data.

In recent work, a number of researchers have examined a variety of ways to deal with this type of data. Their methods span the free recognition of acoustic data using an ASR program and then extracting portions where the ASR output agrees with the quick transcription to using special language models trained on (only or mainly) the QT [1, 2, 3]. Besides being time-consuming, a major problem with using an ASR program to decode the QT data is that errors made by the ASR system during the decoding process will tend to reinforce the same kinds of errors in future. We therefore propose a novel strategy that incorporates the idea of running a decoder on the acoustic signal, but also makes the maximum possible use of any provided QTs during the decoding process to simultaneously increase the accuracy and reduce decoding time. We may succinctly describe our method as one that decodes each speech sample with a special very-high-order language model built from the particular QT for that sample, but modified to account for missing disfluencies or the possibility of incorrect insertions and deletions.

2. Approach

We now give a detailed description of our approach. It is characterized by a rapid alignment of the acoustic signal to specially designed word lattices that allow for the possibility of either skipping erroneously transcribed or untranscribed words in either the transcript or the acoustic signal, and/or the insertion of an optional disfluencies before the onset of every word.

However, flexible alignment is only one of a number of steps (albeit the most important one) involved in preparing the QT for training acoustic models. The full sequence of processing steps is as follows.

Step 1: Generating an initial set of references from the given transcripts. Many of these will be excessively long and often span entire broadcast shows. A good deal of text normalization and filtering happens at this step to canonicalize spellings and automatically correct any obvious transcription errors.

Step 2: Cutting the given waveforms to span exactly the set of references generated in the previous step.

Step 3: Identifying waveforms and references that are too long to be reliably processed (in our case these were segments longer than 30 seconds)

Step 4: Forcibly aligning these long waveforms to their reference transcripts with partial flexibility. In this alignment, we allow both for reference words and portions of the waveform to be skipped as otherwise the alignment would fail on a number of waveforms.

Step 5: Identifying the location of pauses in the aligned output and iteratively cutting the waveforms at increasingly shorter pauses until all waveforms are of manageable length (30s or less).

Step 6: Generating lattices for flexible alignment and flexibly aligning all of the resultant waveforms.

Step 7: Converting the output of the recognizer from the previous step into reference transcripts for training acoustic models.

2.1. Flexible Alignment

The flexible alignment procedure described by us requires the construction of a special recognition search graph (lattice) for each reference transcript. We do this programatically, by processing every transcript to generate a hypothesis search graph that has the following properties.

1: Every word is made optional. This allows for arbitrary amounts of the transcript to be skipped while still entertaining the possibility of resynchronizing with the waveform at a later point.

2: Every word is preceded by either an optional *garbage* word, which we call the @reject@ word, or one of a certain number of disfluencies, namely, um, uh, uhhuh, huh, hmm or uhuh. This allows for arbitrary amounts of the acoustic signal to be skipped while still entertaining the possibility of re-synchronizing with the transcription at a later point. It also allows some of the words frequently omitted in QT to be recovered.

3: Every word is followed by an optional pause of variable length. In our approach the pause word is modeled using a special pause phone that is trained on background noise.

The @reject@ word nominally represents out of vocabulary items in the recognition language model. Consequently it allows for the possibility of unknown words being present in the acoustic signal. These words are matched to the @reject@ word if none of the known words provides a better acoustic match. Because the @reject@ word ought to be able to match arbitrary unknown words, it is composed of a special phone that we call *rej*. This phone is trained on the acoustic data from all of the out of vocabulary (OOV) words. Also, because @reject@ nominally stands for out of vocabulary items, we encode it in the search graph with a probability that is roughly equal to that of an OOV item (the OOV rate) as measured on a development test set. Similarly, the probabilities of transitioning through each of the disfluencies are likewise determined empirically to be the relative frequencies of each disfluency on a development test set.

2.2. Class Based Lattices

To allow for a compact representation of the flexible alignment lattice, we encode the optional elements as class tags in the top-level lattice. These class tags subsequently expand out into sub-lattices of their own. Although arbitrarily nesting such lattices is in theory possible, we obtained the most compact representation at this level where each of the sub-lattices of the top-level lattice contained only terminal symbols. Figure 1 shows a top-level lattice and Figures 2 and 3 show the nested sub-lattices.

Word	Probability
@reject@	0.003
hmm	0.0000293
huh	0.0000162
uh	0.014
uhhuh	0.00015
uhuh	0.00015
um	0.0000162

Table 1: Probabilities of transitioning into each of the arcs of the OPT_NSREJ lattice. These were estimated empirically from held out data. These probabilities are expectedly low because the vocabulary of the ASR system is large.

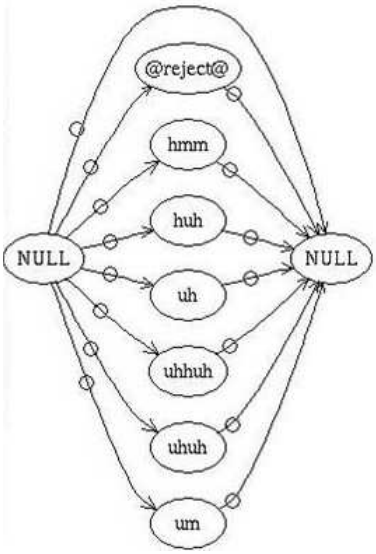

Figure 2: The sub lattice that encodes disfluencies and the @reject@ word. This is represented by the node labeled OPT_NSREJ in the top-level lattice. Transition probabilities are not shown here, but are listed in table 1.

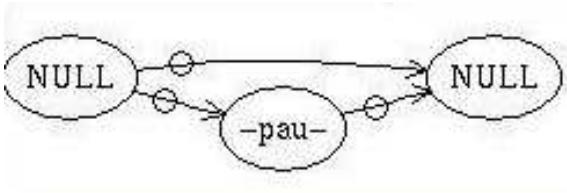

Figure 3: The sub lattice that encodes an optional pause before each word. This is represented by the node labeled OPT_PAU in the top-level lattice. The probability of skipping the pause word was empirically estimated to be 0.6 from the proportion of pause words words in previous recognition outputs.

3. Method

3.1. Data Used

Our empirical tests were based on the the broadcast news (BN) data from the TDT-4 collection released by LDC. The LDC baseline transcripts came from closed captioned television shows. More recently, the LDC has released a manually corrected subset of these transcripts. These were used to upper-bound the improvement that can be obtained with automatic QT cleanup procedures.

As a further point of reference, we also tested TDT-4 transcripts that were automatically generated by Cambridge University's BN recognition system using a biased language model (LM) trained on the closed captions. Their automatic transcriptions, which we call the Cambridge transcripts, were generated by a fast, stripped-down version of their regular ASR system that had the best performance in the NIST 2003 BN STT evaluations [4]. These transcripts have also been used by Cambridge University for their own experiments on lightly-supervised acoustic model training [2].

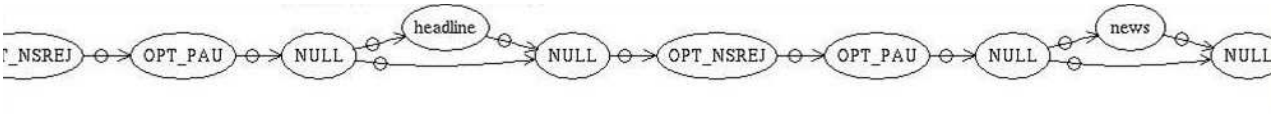

Figure 1: Part of a top level flexible alignment lattice for the reference transcript "this is headline news with judy fortin". Nodes labeled with upper case letters (except NULL) are actually sub-lattices that are dynamically expanded in the recognizer.

The acoustic models used for flexible alignment were trained on the 1996 and 1997 Hub4 Broadcast News acoustic training data. They consist of gender-independent, within-word genonic triphones using standard 39-dimensional MFCC features. The recognition vocabulary was selected to be the top 125000 words from the Hub4 language model training data using the maximum likelihood procedure due to [5] tuned on the TDT-4 data set on which the resultant vocabulary was found to have a negligible out of vocabulary (OOV) rate.

3.2. Evaluation strategy

The final output of our flexible alignment procedure is a set of reference transcripts that is (hopefully) of better quality than the original closed captioned transcripts that were distributed by LDC. In order to measure the quality of these transcripts, we decided to train acoustic models with them and evaluate the performance of these acoustic models on three different evaluation data sets. Approximately 36 hours of these transcripts were discarded as unsuitable for training. 14 of these hours were dicarded by the alignment procedure itself as unalignable data. The remaining 22 hours were discarded post-hoc because the proportion (20%) of @reject@ words in them exceeded our intuitive threshold for being acceptable.

The three data sets used to evaluate the acoustic models were the 2003 and 2004 TDT-4 development test sets defined by the DARPA EARS program participants (denoted **dev2003** and **dev2004**) and the RT-03 Speech To Text (STT) evaluation data set for broadcast news distributed by NIST (denoted **eval2003**).

Besides the TDT-4 reference and acoustic data, the data used for acoustic model training includes 1996 and 1997 Hub4 English Broadcast News Speech (75 hours and 71 hours, respectively). The acoustic training data was processed with cepstral normalizations and vocal tract length normalization (VTLN). 52 dimensional Mel frequency cepstrum (MFC) features (13 MFCcs + first, second and third order differences) are reduced to 39 dimensions using Heteroscedastic Linear Discriminant Analysis (HLDA). The acoustic models were trained as follows: after training phonetically tied mixture (PTM) models, the models were clustered and genonic acoustic models were trained. Phonetic models had the usual three-state HMM structure with left-to-right transitions and self-loops (enforcing a minimum duration of 3 frames). All models used were trained using the maximum likelihood criterion [4].

4. Results and Discussion

4.1. Word Error Rates on Hub-4 Broadcast News Data

We now report the results of evaluating the various sources of TDT-4 transcripts. Evaluation was performed indirectly by adding in TDT-4 transcripts from each source into the acoustic model training procedure and observing the ASR accuracy of the resultant models. All the models were trained on the same 146 hours of Hub4 Broadcast news training data from 1996 and 1997, but with the different supplemental sources of TDT-4 data, and in the baseline, no supplemental TDT-4 data. The versions of the TDT-4 transcripts that we evaluated include the original and hand-corrected closed captions from the LDC, Cambridge University's ASR transcripts, and our own flexibly aligned transcripts. Transcripts that did not align during training were simply discarded. We also discarded any training shows from the two-week period from which the two development test sets were drawn and shows that didn't belong to the subset of hand-corrected TDT-4 transcripts provided by the LDC.

Our ASR procedure included the following steps: Segmentation, computation of cepstra, removal of cepstral mean, normalization of variances, application of vocal tract length normalization (VTLN), HTK lattice generation using a within-word acoustic model and a bigram language model, and finally rescoring of HTK bigram lattices using a 5-gram almost-parsing language model [4]. In Table 2, the initial WERs (%) refer to the 1-best WER of the bigram HTK lattices and the final WERs (%) refer to the WER after 5-gram language model rescoring.

To reduce the influence of varying segmentation strategies between the systems employed by Cambridge University and us, we trained two variations of acoustic models from the Cambridge transcripts. One used their own segmentations and the other had segment lengths determined by our waveform segments. Consequently, there were five acoustic models evaluated on the three test sets described in Section 3, as shown in Table 2.

As Table 2 shows, the Flexalign model produced the lowest word error rate (WER) after first-pass decoding on both the Hub4 Broadcast News 2003 and 2004 TDT-4 development test set and the lowest WER on all test sets after rescoring using the 5-gram almost-parsing language model. On the Eval 2003 test set, the performance of the Flexalign model is still competitive with the performance of the two best models. Taking into account that the flexible alignment approach is faster than real time, these results verify that this approach is at once effective and efficient.

4.2. Varying the amount of data

Besides comparing the quality of transcripts generated using different methods for acoustic model training, we were also interested in investigating the effect of varying the amount of training data for our proposed approach. Table 3 presents the WERs using a similar evaluation framework to Table 2. The Flexalign-subset model was trained using Flexibly aligned TDT-4 transcripts from the same subset of shows as in the LDC-hand-corrected data (325 hours). The Flexalign-full model was trained using all available Flexibly aligned transcripts (366 hours). For each model, As can be seen from Table 3, increasing the amount of training data also scales up the performance of the the flexible alignment procedure.

Model	WER on dev2003		WER on eval2003		WER on dev2004	
	Initial	Final	Initial	Final	Initial	Final
Baseline	22.0	17.8	19.9	14.9		
LDC-raw	21.3	16.8	19.7	14.7	22.4	18.9
LDC-hand-corrected	20.9	15.9	18.9	13.9	22.0	18.1
CUED-CUED-segs	21.1	16.0	19.2	14.1	21.9	18.2
CUED-SRI-segs	20.7	15.9	18.9	14.0	22.0	18.2
Flexalign	20.8	15.8	19.3	14.4	21.8	18.0

Table 2: WER results (%) using models trained with flexible alignment transcripts compared with those with (1) only Hub4 transcripts (no TDT-4 data) (2) Raw closed captioned transcripts processed just as with our flexible alignment procedure except that optional words were not used and unalignable transcripts were discarded (3) A hand-corrected subset of (2) provided by LDC (4) Transcripts from Cambridge (CUED) with lengths determined by CUED's segmentation and (5) Transcripts from Cambridge but with lengths determined by our waveform segments. The Initial WERs refers to the 1-best WER of the bigram HTK lattices and the final WERs refer to the WER after 5-gram language model rescoring.

Model	WER on dev2003		WER on eval2003		WER on dev2004	
	Initial	Final	Initial	Final	Initial	Final
Flexalign-subset	20.8	15.8	19.3	14.4	21.8	18.0
Flexalign-fullset	20.9	15.8	19.0	14.2	21.8	17.7

Table 3: WER results using models trained with varying amounts of flexibly aligned transcripts.

4.3. Precision and Recall on Disfluency Insertion

A final experiment examined the accuracy of inserting disfluencies from the flexible alignment approach. The experiment was performed on the 1996 Hub4 Broadcast News training transcripts by initially removing all of the disfluencies from the set and re-inserting them using the flexible alignment approach. The original transcripts with disfluencies served as the reference set. We then evaluated the precision (proportion of inserted disfluencies that were correct) and the recall (proportion of correct disfluencies that were inserted). On the 1996 broadcast news acoustic training data, the flexible alignment approach obtains a precision of 68% and a recall of 54%. While it is hard to assess these numbers in absolute terms for lack of a point of reference, we take them as further indication that the flexible alignment approach works reasonably on its intended task of fixing inaccuracies in quick transcriptions.

5. Conclusions

We have shown a fast and effective approach for correcting (some of) the errors found in quick transcripts (e.g., closed captions) of speech, thus making them more suitable for training acoustic models. We obtained consistently better accuracies on three BN testsets from the flexibly aligned transcripts than with the originals, and results that are comparable to or better than those with transcripts generated by free recognition, which are more time-consuming to produce. In future work, we hope to refine this procedure by employing more sophisticated flexible alignment lattices that use higher-order context dependent transition probabilities.

6. Acknowledgments

This research was supported by DARPA under contract MDA972-02-C-0038. Distribution is unlimited. We are grateful to Phil Woodland and Cambridge collaborators for making their ASR transcripts available for our work, as well as for useful discussions.

7. References

[1] L. Lamel, J.-L. Gauvain, and G. Adda, "Lightly supervised acoustic model training", *in ASR2000—Automatic Speech Recognition: Challenges for the new Millenium*, Paris, Sep. 2000.

[2] H. Y. Chan and P. Woodland, "Improving broadcast news transcription by lightly supervised discriminative training", *in Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, Montreal, May 2004, To appear.

[3] L. Nguyen and B. Xiang, "Light supervision in acoustic model training", *in Proceedings of the IEEE Conference on Acoustics, Speech, and Signal Processing*, Montreal, May 2004, To appear.

[4] A. Stolcke, H. Franco, R. Gadde, M. Graciarena, K. Precoda, A. Venkataraman, D. Vergyri, W. Wang, J. Zheng, Y. Huang, B. Peskin, I. Bulyko, M. Ostendorf, and K. Kirchhoff, "Speech-to-text research at SRI-ICSI-UW", *in DARPA RT-03 Workshop*, Boston, May 2003, http://www.nist.gov/speech/tests/rt/rt2003/spring/presentations/sri+-rt03-stt.pdf.

[5] A. Venkataraman and W. Wang, "Techniques for effective vocabulary selection", *in Proceedings of the 8th European conference on Speech Communication and Technology*, pp. 245–248, Geneva, 2003.