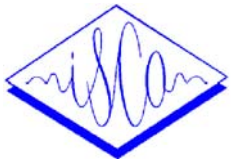

6th European Conference on
Speech Communication and Technology
(EUROSPEECH'99)
Budapest, Hungary, September 5-9, 1999

AUTOMATIC VERIFICATION OF BROADCAST NEWS TRANSCRIPTIONS

Michael Pitz, Sirko Molau

Lehrstuhl für Informatik VI, RWTH Aachen – University of Technology
Ahornstraße 55, 52056 Aachen, Germany
{pitz,molau}@informatik.rwth-aachen.de

10.21437/Eurospeech.1999-173x

ABSTRACT

In this paper we present a method for automatically detecting erroneous training scripts for speech corpora like Broadcast News and Switchboard. Based on the Hub-4 task we will report on the performance of error detection with the proposed method and investigate the effects of both manually and automatically cleaned training corpora on the performance of the RWTH speech recognition system. Our approach uses a forced Viterbi alignment on the training data and evaluates different transcription quality measures. The following three criteria proved to be useful to automatically detect most transcription errors:

- the difference between the final Viterbi alignment HMM state and the last state according to the transcriptions
- the normalized acoustic word scores
- the location of the boundary between adjacent segments obtained by forced alignment

With manually corrected scripts we achieved a WER reduction on the 1996 HUB-4 eval. corpus. The recognizer's performance improved mainly on clean planned speech segments. Whereas the improvements were minor on these hand-transcribed training data, automatic training script verification will become more important for automatically transcribed new speech corpora.

1. INTRODUCTION

In the last few years, the focus in ASR research has shifted from the recognition of clean planned speech (i.e. WSJ) to the more challenging task of transcribing found speech like broadcast news (Hub-4 task) and telephone conversations (Switchboard). Available training corpora tend to become larger and more erroneous than before, as transcribing found speech is more difficult. The importance of transcription verification was highlighted by the 1997 Hub-4 Broadcast News evaluation. A number of participating sites reported efforts to clean transcriptions of the training material hereby improving the quality of their acoustic models [1, 2]. Either the whole corpus was manually checked and corrected, or suspicious speech segments with bad acoustic scores were rejected during training. Our first tests with the RWTH large vocabulary speech recognition (LVCSR) system [3] on the 1996 Hub-4 evaluation task supported this procedure. We obtained a reduction in WER (table 1) when training on a subset of manually checked 46 hours compared to 76 hours of original data as released 1996 and 1997 by LDC. Even though the subset contained 40% less training data, the WER decreased. This indicated that the system was rather sensitive to incorrect transcriptions.

We then carried out a number of further test on the Broadcast News speech corpus to investigate the effects of erroneous training scripts in more detail. The work focussed on two main questions:

- How to detect transcriptions errors in the training data automatically?
- Which improvements can be obtained by manually correcting training scripts? Will minor transcription errors degrade the performance of the recognizer or will they make the acoustic models more robust?

1996/97 Hub-4 training corpus	size	WER
manually checked	46h	36.7%
complete	76h	37.1%

Table 1: Recognition results on Hub-4 '96 evaluation test set, obtained with a *preliminary* RWTH system trained on different training corpora. All WER reported in this paper are obtained by NIST scoring.

2. AUTOMATIC SCRIPT VERIFICATION

2.1. Verification algorithm

There are two different types of errors in the Boradcast News training corpus:

- incorrect transcriptions, i.e. wrongly transcribed words or missing words
- incorrect segment boundaries, i.e. incorrect begin or end times of speech segments

Our approach to detecting these errors is based on a forced Viterbi alignment on the training data and the evaluation of different transcription quality measures for each speech segment.

First, low resolution acoustic models (2000 tied states, 60k Gaussian densities with pooled variances, gender independent models) were trained on 46 hours of manually cleaned Hub-4 training data. The alignment was then carried out with the RWTH speech recognizer, an HMM-based LVCSR system with decision tree clustered acoustic models of continuous Gaussian mixture densities.

2.2. Transcription quality measures

We investigated six criteria to automatically detect erroneous scripts. Each training segment was classified according to

(1) whether or not the optimal path in the dynamic programming (DP) time alignment reached the terminal HMM state,

(2) the width of the beam required for the alignment,

(3) the acoustic sentence score, normalized to the number of time frames,

(4) the normalized acoustic word scores, and

(5) the duration of each word in the segment.

In addition, adjacent training segments were joined to compare the boundary given in the training script with it's location according to the forced alignment (6).

Segments were sorted by quality according to these measures in order to inspect the worst ones first.

2.3. Evaluation of quality measures

A preliminary check of the segment quality measures proved that the criteria (1), (4), and (6) were useful in detecting script errors. On the contrary, there was only little correlation between faulty transcriptions and the criteria (2), (3), and (5).

Segment-wise criteria, (1)–(3): Measure (1) mainly detected major errors in training scripts like whole sentences that were missing in the transcriptions or incorrect segment boundaries. Measures (2) and (3), however, were highly speaker- and focus condition dependent (table 2) and therefore of little use in detecting script errors.

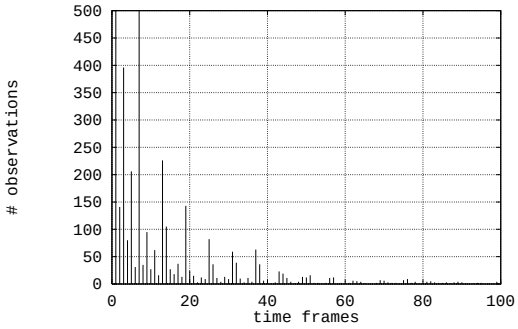

Figure 1: Histogram over differences between the final HMM state in the DP alignment and the terminal state according to the training script.

Incorrectly transcribed single words were not detected by any of these segment-wise criteria due to the usually long training segments. The acoustic sentence score of a given segment is the normalized sum of acoustic word scores. Hence, the poor score of one wrongly transcribed word may be masked by the scores from the other words. Equally, the DP algorithm may reach the terminal HMM state even if there is a transcription error at the beginning or middle of a segment.

condition	description
F0	baseline broadcast speech
F1	spontaneous broadcast speech
F2	speech over telephone channels
F3	speech in the presence of background music
F4	speech under degraded acoustical conditions
F5	speech from non-native speakers
FX	all other conditions

Table 2: Focus conditions in the Broadcast News speech corpus.

Word-wise criteria, (4) and (5): Criterion (4) correctly indicated missing or wrongly transcribed single words, but also utterances with strong background noise or overlapping speech. On the contrary, measure (5) gave only little evidence of script errors as the duration of words is basically speaker- and context dependent. Words with significantly shorter duration than their average were rarely observed, which could have been caused by the minimum word length constraint of our 6-state HMM topology.

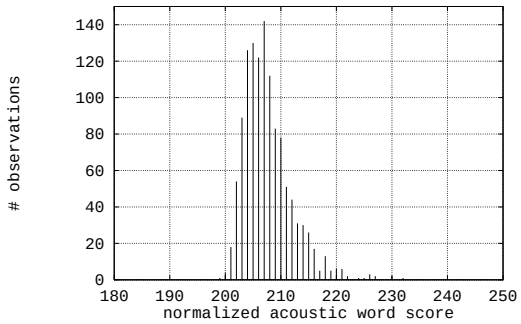

Figure 2: Histogram over normalized acoustic word scores for the word 'PRESIDENT'.

Finally, the **across-segment criterion (6)** indicated wrong segment boundaries as well as major transcription errors similar to quality measure (1). An error was only suspected if the segment started later according to the transcription, or if it ended earlier. The other two cases were usually caused by silence frames at the segment begin or end.

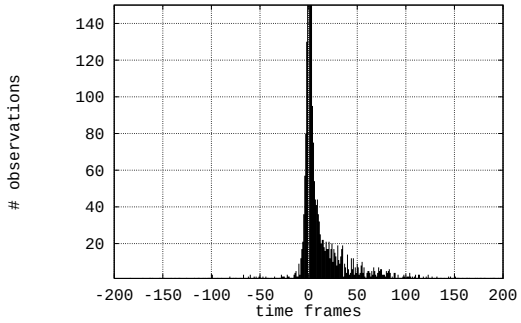

Figure 3: Histogram over time differences between the segment boundaries of adjacent segments. Displayed are critical positive time differences at the begin of segments and negative at their ends.

As (2), (3), and (5) showed poor efficiency in detecting transcription errors, they were excluded from further analysis.

2.4. Application of quality measures

Starting from a forced Viterbi alignment we calculated the difference Δ between the final HMM state and the terminal state according to the script for each segment (figure 1). Likewise, we calculated the normalized acoustic score s_w for each word in the segment (figure 2), and the time difference Δt between the segment boundaries of adjacent segments according to (6) (figure 3). We then computed the mean score $\bar{s}_w$, variance σ_w, and the number of observations N_w for each word as well as the overall mean $\bar{s}_{all}$ and variance σ_{all} of all normalized word scores.

A segment was considered to have potential script errors if

(1) $\Delta > 10$,

(4) $N_w > 10 : (s_w - \bar{s}_w)/\bar{s}_w > 3\,\sigma_w$
$N_w \leq 10 : (s_w - \bar{s}_{all})/\bar{s}_{all} > 3\,\sigma_{all}$,

and / or

(6) segment begin: $\Delta t > 500$ ms,
segment end: $\Delta t < -500$ ms.

That is, if a word did not occur frequently enough in the training corpus ($N_w \leq 10$) we used the overall mean word score $\bar{s}_{all}$ and variance σ_{all} as fallback values.

The deviations were considered to be significant only if they exceeded the given values. The thresholds were chosen in such a way that about one third of the corpus was marked. This was the order of suspicious segments reported by other groups.

3. RESULTS

3.1. Segment classification statistics

We applied the described method to the full 1996/97 Hub-4 training corpus (15 389 segments, 76 h). 35% of the corpus (5 429 segments, 28h) was marked as possibly erroneous. Most of these segments (72%) were tagged because they contained words with bad acoustic scores (4). Criteria (6) and (1) supplied 13% and 7% of the bad segments, respectively. The remaining 8% were classified as bad according to two or all three criteria.

The marked segments were manually corrected afterwards. During the correction process we estimated that the rate of false alarms was in the order of 25%, which means that most segments labelled as 'bad' actually contained wrong transcriptions or segment boundaries. After correcting, 75 hours of training material remained; only one hour worth of data was considered to be too bad for training because of overlapping or unclear speech.

In order to investigate the number of errors that remained undetected we first examined a sample of 25% of the segments from training CD 4 which were not tagged before. From these segments, only 11% contained errors which were not detected by our method.

Later we analysed another 3.5 hour subset (training CD 1 with 1 352 segments) of the training corpus in more detail. All segments of this subset were manually checked. Transcription errors were corrected and scored according to four categories: minor, medium, and major script error, and too bad for training.

When evaluating the performance of our quality measures we focused on the last three categories. Segments falling into the category 'minor error' contained untranscribed noise items or errors affecting only single phonemes like *get* ↔ *got*.

Table 3 shows detailed results for the different quality labels.

error type	# manually selected segments	# automatically detected segments	
	abs.	abs.	rel.
minor	192	62	32%
medium	72	30	42%
major	20	12	60%
too bad	209	48	23%

Table 3: Statistics of automatic error detection on CD 1

These results do not confirm the impression we had when correcting the automatically tagged segments. They also contradict the findings from examining the subset of CD 4. Reviewing the data there were a number of facts that made us believe that this particular CD is not representative for the whole corpus:

- On CD 1 the rate of false alarms was 52%, significantly higher than the average (about 25%).
- The 48 segments automatically labelled as 'too bad' make up 20% of this category for the 1996/97 training corpus (233 automatically labelled 'too bad' segments), although CD 1 contains less than 5% of the whole corpus.
- From the 1 352 segments our method marked only 286 (21%) as possibly errorneous, which was well below the average percentage of tagged segments in the whole corpus (35%).

A possible explanation could be that the data on CD 1 are of especially bad quality for some reason. It might have been better not to compute word score statistics for the whole corpus but rather per CD or even per show or speaker.

3.2. Effects on the word error rate (WER)

The transcription verification approach presented here was further checked by evaluation tests with acoustic models obtained from different training data sets:

- the complete 1996/97 Hub-4 training corpus which amounts to about 76 hours of speech data,
- the manually checked 46 hour subset, in which all incorrect segments were rejected and only a few obvious errors were corrected, and
- the 75 hour subset, where an overall of 22 hours of erroneous segments were automatically detected and manually corrected thereafter.

All recognition tests were carried out with a single pass integrated trigram Viterbi decoder based on word-internal triphones.

While our preliminary Hub-4 system performed better when trained on clean but less data (table 1), we observed a different behaviour of the system that was optimized for this task (table 4).

1996/97 Hub-4 training corpus	size	WER
manually checked	46h	33.6%
complete corpus	76h	32.8%
automatically checked + manually corrected	75h	32.5%

Table 4: Recognition results on Hub-4 '96 evaluation test set, obtained with our LVCSR system *optimized* for the Hub-4 task using different training corpora.

A closer inspection of the recognition results revealed that the extensive manual transcription correction gave significant improvements for planned clean speech (F0 condition, table 2) only. The word error rate remained almost constant in the other focus conditions.

Acoustic models trained on the manually checked 46 hour subset performed well under the F0 condition, too. However, the smaller amount of training data made them less robust for more difficult conditions (spontaneous speech, background noise), which is why they were outperformed by acoustic models trained on the complete uncorrected corpus.

4. CONCLUSION

The transcription verification method proposed in this paper is able to detect major transcription errors. Three of the six investigated quality measures proved to be useful to tag faulty scripts. In about half of all cases the criteria even marked the position of the error correctly, at least within the range of a few words. We expect an improved performance by adjusting the thresholds for segment classification, which have not been optimized so far. Furthermore we intend to combine the quality criteria described here with confidence measures, which have been shown to reduce tagging error rates on different corpora [4].

From the point of view of the WER of our optimized Hub-4 speech recognition system, the question of quality vs. quantity of hand-transcribed training data is not easy to answer. The recognizer performed better on the difficult Hub-4 '96 evaluation test set when trained on more but unclean data. The overall improvement obtained by extensive manual corrections was relatively small. Only segments of clean planned speech gained significantly from these efforts.

With further increase of training corpora size in future, manual correction will become infeasible. The main goal will then be to accept or reject suspicious segments or even parts of them rather than manually correcting the scripts.

Finally, the importance of verification methods will increase when using automatically transcribed training corpora like TDT-2 with 800 hours of speech data. The transcriptions of such corpora will have a significantly higher error rate than manually transcribed ones. In addition, the error types may differ from what has been observed so far. Both will affect the performance of LVCSR systems and our verification approach. Thus, the automatic transcription verification will remain a challenging task in future.

Acknowledgements
This work is part of a joint effort of the University of Technology (RWTH) and Philips Research Laboratories Aachen, Germany.

5. REFERENCES

[1] J.L. Gauvain, L. Lamel, G. Adda, "The LIMSI 1997 Hub-4E Transcription System", *Proc. DARPA Speech Recognition Workshop*, Lansdowne, VA, Feb. 1998.

[2] P.C. Woodland, T. Hain, S.E. Johnson, T.R. Niesler, A.Tuerk, E.W.D. Whittaker, S.J. Young , "The 1997 HTK Broadcast News Transcription System", *Proc. DARPA Speech Recognition Workshop*, Lansdowne, VA, Feb. 1998.

[3] H. Ney, L. Welling, S. Ortmanns, K. Beulen, F. Wessel: "The RWTH Large Vocabulary Continuous Speech Recognition System", *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Seattle, USA, pp. 853-856, May 1998.

[4] F. Wessel, K. Macherey, R. Schlüter: "Using Word Probabilities as Confidence Measures", *Proc. IEEE International Conference on Acoustics, Speech and Signal Processing*, Seattle, USA, pp. 225-228, May 1998.