Evaluation of Interactive User Corrections for Lecture Transcription

Henrich Kolkhorst, Kevin Kilgour, Sebastian Stüker, and Alex Waibel

International Center for Advanced Communication Technologies – InterACT
Institute for Anthropomatics
Karlsruhe Institute of Technology, Germany

henrich.kolkhorst@student.kit.edu,
{kevin.kilgour, sebastian.stueker, alexander.waibel}@kit.edu

Abstract

In this work, we present and evaluate the usage of an interactive web interface for browsing and correcting lecture transcripts. An experiment performed with potential users without transcription experience provides us with a set of example corrections.

On German lecture data, user corrections greatly improve the comprehensibility of the transcripts, yet only reduce the WER to 22%. The precision of user edits is relatively low at 77% and errors in inflection, case and compounds were rarely corrected. Nevertheless, characteristic lecture data errors, such as highly specific terms, were typically corrected, providing valuable additional information.
Index Terms: speech recognition, user study, transcript correction, lectures

1. Introduction

Recording and archiving of lectures is feasible and practiced at several universities (e.g., the MIT OpenCourseWare project [1]). Nevertheless, automatic speech recognition (ASR) on lecture data is non-trivial, for example due to highly specific contents and spontaneity in speech style. Since word error rates (WERs) should be less than 25% for a lecture archive to be perceived useful [2], careful adaptation is needed.

Besides enabling searchable archives of lectures, ASR is necessary for spoken language translation. At KIT, significant research is conducted to enable simultaneous translation of lectures [3], requiring good speech-to-text performance for further processing. Misrecognition of content words, such as substituting the word "censor" for "tensor" in a mathematical context, impairs the readability of transcripts and can cause substantial errors in subsequent computation steps.

Most of these errors can easily be corrected by humans. However, professional transcription on a larger scale is unrealistic due to required time and resulting costs. Especially large-vocabulary recognizers often contain the correct words in their lattice (e.g., a 1-best WER of 55% on lecture data compared to a lattice WER of 30% [4]) and, given adequate tools, users of lecture archives can quickly correct such errors. Ideally, corrections of existing transcripts should also be used to improve future recognition results on similar data.

In this work, we investigate the quality of error corrections by users of a lecture archive and the usability of such corrections for system adaptation. We present an interface for browsing transcriptions in which error corrections can be made quickly, along with the results of a user experiment involving the correction of German lectures.

After giving an overview of related work on user corrections and their utilization for adaptation in Section 2, we describe the interface and experimental setup in Section 3 and 4. Results of a user experiment are presented and analyzed in Section 5.

2. Related Work

In a setting of webcast archives, Munteanu et al. describe a "wiki-like" transcript edit tools for lectures, which can be used to correct errors in speech recognition output [5]. In a user study, students corrected lectures (from a single course), reducing the WER of the ASR output by 53%. However, the initial WER of 50% to 60% was quite high. If the actual transcription of a sixth of a lecture is available, transformation-based learning from the correct transcript has been shown to reduce WER by 12.9% [6].

Within the framework of the MIT OpenCourseWare project, Hsu and Glass investigate the possible improvements based on partial user transcriptions by adapting language model (LM) interpolation parameters. They show that with 300 words of transcription, adaptation on recognizer hypotheses is outperformed and about 1% absolute reduction of WER can be achieved from a 33.2% baseline. However, they use parts of the reference transcription and not actual user data.

Yu worked on correction of MIT lecture transcripts based on re-recognition of error-prone regions [7]. Using oracle corrections from reference transcript, relative WER reductions of 39% were obtained. In a user test, precision and recall of user corrections were both at 97%, but no re-recognition was performed with actual user data. The correctors in the test were primarily speech researchers and therefore probably aware of transcription guidelines.

Ogata and Goto used confusion network output for error correction during online speech recognition and showed that in theory, 83% to 99% of errors in podcast transcripts could be corrected based on confusion networks [8]. Additionally, Ogata et al. investigated user corrections in their "PodCastle" podcast transcription service [9, 10]. User corrections reduced WER by more than 50% and 46 hours of corrected podcast data were collected. However, the actual correction data was not analyzed in detail.

Based on user correction, Ogata et al. used Maximum Likelihood Linear Regression (MLLR) and subsequent Maximum A Posteriori Estimation (MAP) to adapt the acoustical model. The model adaptations yielded relative improvements in WER up to 23% when a large number of episodes had been corrected (between 7 and 20 hours of training data) [9].

Recent work has investigated the use of platforms for human intelligence tasks, such as Amazon's Mechanical Turk (MTurk), for transcription and correction of ASR transcripts. Marge et al. used MTurk workers to transcribe clean instructional audio segments and found the quality of transcripts to be at 5% WER. Considering cost and accuracy, they suggest using three to five workers for transcrip-

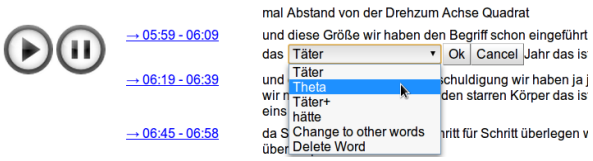

Figure 1: Screenshot of alternate confusion network hypotheses

tion [11]. Lee and Glass use a two-stage process to generate lecture transcripts from MTurk tasks. As a first step, short utterances are transcribed, yielding a WER of 16%. In a second stage, workers are asked to correct a baseline transcription from the first stage. Integrating a detection for poor quality transcripts and giving workers performance estimations as feedback, they report a WER of 10.2% after the second stage [12].

3. Interface Features

To facilitate the interaction with lecture archives, interface usability and familiarity is essential, making web applications a natural choice. The interface has been implemented solely based on HTML5 standards without the need for browser plugins to assure platform independence. The Google Web Toolkit[1] was used for implementation.

ASR lattices are converted to *confusion networks* [13], a representation with total ordering of word hypotheses which are collapsed into "clusters" at specific time slots. This enables the display of time-aligned alternate hypotheses in the interface. Playback of the lecture's audio recording can be started from the beginning or users can jump to specific utterances. By default, the 1-best transcript is displayed and the current utterance is subtly highlighted during audio playback.

By clicking on a word, a list of alternate hypotheses for the time slot along with the option to delete or enter a different word is displayed (Figure 1). To prevent a cluttered or complex interface, users cannot move words between utterances or insert words at specific time slots. Instead, existing slots can be modified to consist of multiple words. Changes are saved instantaneously to enable frictionless interaction. It is possible to redirect hypotheses of online recognition into the web interface.

4. Experimental Setup

A user study was performed to evaluate the correction performance of students using the web interface. Since corrections will typically take place "offline" (not during the lecture), the initial ASR hypotheses have been generated by a system combination to achieve a high-quality basis for subsequent editing.

4.1. Corpus Characteristics

For the experiment, German lectures from a variety of topics were used. The lectures form a subset of the KIT Lecture Corpus for Speech Translation [14]. The lectures "Algorithms for Planar Graphs" (ALGO), "Formal Systems" (FORM), "Cognitive Systems" (COGSYS), "Machine Translation" (MT) and "Multiprocessors" (PROC) cover different areas of computer science. The "Technical Mechanics" (MECH) lecture is from an unrelated, but still technical area, whereas the lectures about "Population Geography" (GEO), "World War 2" (WW2) and "Copyright Law" (LAW) cover non-technical topics.

[1] https://developers.google.com/web-toolkit/

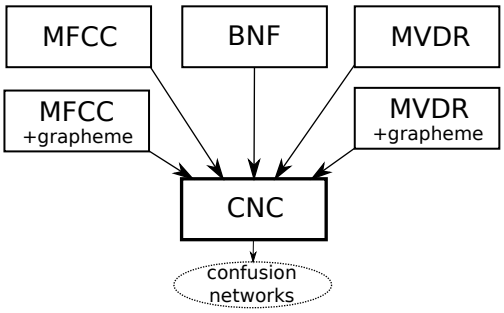

Figure 2: Decoding systems for generation of baseline transcription

The lectures were recorded with a close-talking microphone and the audio data has a sampling rate of 16kHz. The speaker style varies significantly. Some lectures contain many hesitations (COGSYS) whereas others are characterized by false starts (MT) or significant amounts of read formulas (MECH, ALGO).

The lectures have been divided into sections edited by the users (EDIT set) and unedited sections used for evaluating adaptation performance (EVAL set). Only lectures in which the unedited sections contained more than 1000 words have been included in the EVAL set to make it reasonably different from the EDIT set.

4.2. Baseline ASR System

The baseline hypothesis which are displayed in the web interface and are editable by the user are produced with the Janus Speech Recognition Toolkit's Ibis Decoder [15] through a confusion network combination (CNC) [13] of five speaker independent systems developed for the 2011 Quaero Evaluation as depicted in Figure 2. It is an improved version of the 2010 evaluation system [16] and similar to the Spanish system described by Kilgour et al. [17]. The underlying systems use three different frontends, mel-frequency cepstral coefficients (MFCC), warped minimum variance distortionless response (MVDR), and MVDR based bottleneck features (MVDR-BNF). Additionally, two systems use graphemes instead of phonemes. The system combination has been chosen to provide state-of-the-art transcripts as basis for corrections. The language model is built from the transcripts of the quaero training data, scraped newspaper data and webdumps.

The vocabulary is case-sensitive and fairly large containing roughly 300k sub-words and 480k pronunciation variants. The sub-words are used in order to improve the recognition of compound words. Sub-words of the 1-best hypothesis are merged appropriately for display in the user interface. A substantial amount of sub-words are, however, also full words. The sub word LM doesn't correctly merge all compound words, so many of them are still falsely recognized as multiple independent words.

The word error rates of this setup on the different lectures are listed in the second column of Table 1. Generally, many errors can be attributed to a mismatch between the training data, which consists primarily of broadcast news, and the lecture data. Especially the frequent use of rare and non-German terms causes problems. For example, the term "phrase alignment" is central to the MT lecture, but the lack of the English pronunciation of "phrase" leads to continuous misrecognition and makes many utterances difficult to comprehend.

Lecture	WER ASR	WER User	Rel. WER Improvement (content words only)	fraction of words edited	edit precision	#words	# users
MECH	32.65	21.15	35.2% (30.8%)	17.3%	80.1%	1493	3
GEO	22.85	19.18	16.1% (15.7%)	13.0%	82.9%	1393	3
WW2	28.57	24.96	12.6% (19.0%)	13.1%	71.4%	1019	2
ALGO	35.92	24.76	31.1% (38.9%)	20.3%	68.6%	1836	4
FORM	29.14	20.68	29.0% (34.6%)	17.4%	85.2%	3137	5
COGSYS	33.61	17.39	48.3% (46.6%)	21.1%	89.7%	876	2
MT	38.65	22.98	40.4% (48.7%)	24.5%	83.5%	4704	6
PROC	35.89	26.19	27.0% (28.1%)	19.9%	72.0%	1365	4
LAW	28.73	20.53	28.5% (24.2%)	19.7%	72.4%	2461	4
total	32.71	22.08	32.5% (34.7%)	19.6%	77.4%	18284	11
mean	31.78	21.98	29.8% (31.9%)	18.5%	77.2%	2032	3.7
std. dev.	4.89	2.93	11.1 (11.5)	3.7	7.1	1229	1.3

Table 1: Overview of corrections by lecture (EDIT set). User values are aggregated over all users who edited the particular lecture. Word error rate and precision are case-insensitive. The relative improvement of the WER by user edits is given on all words and on "content words" only (excluding the 1000 most frequent words in training).

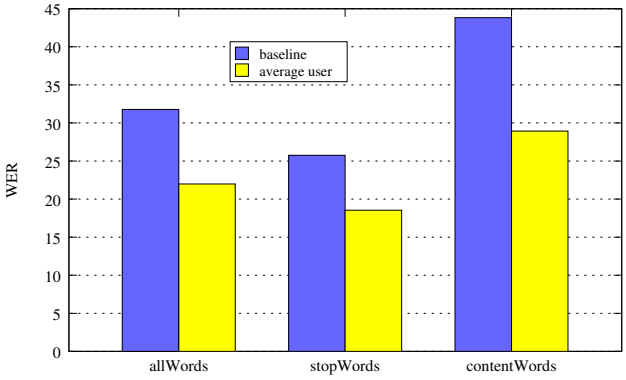

Figure 3: Differences between word error rates of "stop words" (the 1000 most frequent words in the training data) and content words

4.3. Setup of User Study

The experiment was carried out with 11 test subjects, most of them university students with a technical major and some familiarity with the subject matter of the lectures. However, none of them had experience in transcription nor did anybody use the interface before. Each user was asked to correct three lecture segments of five minutes each.

Test subjects were asked to correct the ASR transcripts based on their own judgment, i.e. correct the transcripts primarily to improve readability and correct only the errors that they feel to be problematic. The omission of fine-grained transcription guideline aims to simulate an every-day usage of a learning environment where users will only correct certain subsets of errors and not adhere to detailed rules for editing.

In order to be able to analyze the influence of different correctors and observe the familiarization with and usage of the interface, the experiment was carried out in a controlled lab environment. Lecture segments used in the experiment were edited by at least two subjects and each subject edited segments with, on average, 1669 words of which 393 words were corrected (see Table 2). To enable comparison of different users, all were presented with the unedited recognizer hypothesis.

5. Results of User Corrections

User corrections improved the transcript quality substantially, yet not comparable with professional transcription. The initial (case-insensitive) word error rates of the baseline transcript ranged from 22.9 to 38.7, with technical lectures generally having more errors. In total, users improved the word error rate by about a quarter from 32.71 to 22.08 (cf. Table 1).

On average, users edited every fifth word slot in the baseline hypotheses. However, in almost a quarter of these edits, incorrect edit operations are made. Furthermore, if alternate hypotheses from the confusion network are chosen, the precision drops below 50% (Table 2). This is primarily due to the selection of compound fragments instead of entering the complete word.

Whereas the quality of the baseline transcripts varies substantially, the user corrections reduced the variance of errors, attenuating the difference between technical and non-technical lectures. If errors are analyzed on "stop words" (the 1000 most frequent words in the training data) and "content words" (the rest) separately, the correction performance varies substantially between lectures, with on average slightly greater improvement on content words (see Figure 3).

5.1. Characteristics of User Edits

Based on manual inspection, the user edits substantially increased the readability and comprehensibility of the lecture transcripts due to the correction of words central to the lecture excerpts. However, there are some peculiarities in the user edits which contribute to their relatively low precision.

Spelling errors are relatively common in the user corrections, especially in rare terms. For example, not all users were familiar with the word "tensor" and users frequently used misspelled words like "aligment" in their corrections which is acceptable for human users, but obviously hurts the WER. Similar problems occur, if lecturers use German inflection on English words such as "pointe" as the plural form of "point".

Another frequent issue in user corrections was the treatment of German compounds. Despite the existence of sub-words in the ASR vocabulary, many compounds are still misrecognized in the ASR hypothesis, for example when only one part is correctly detected. Often, users corrected only the first part (replaced it with the actual compound), but did not remove the second part. Since the deletion function was generally used in other cases, it can be assumed that these duplications did not substantially bother the users.

	mean	std. dev.
number of words edited	393.4	96.1
word error rate (case-independent)	21.95	2.80
insertions (%)	8.1	1.3
deletions (%)	34.9	5.3
substitutions (%)	47.0	6.2
word error rate (case-dependent)	24.33	2.96
ratio of edits chosen from confusion network alternatives in relation to total edits (%)	64.6	8.9
precision of edits chosen from confusion networks (case-independent) (%)	46.6	3.0

Table 2: Detailed statistics of user edits, aggregated over all users.
Users select confusion network alternatives in a majority of cases which are mostly incorrect and lead to a high substitution error.

Errors in case were rarely corrected by users. In their own corrections, some users used correct capitalization rules whereas others preferred lower-case corrections. Generally, alternate confusion network hypotheses were chosen regardless of correct capitalization. Ignoring the case, if multiple users corrected the same hypotheses, the inter-user agreement of 89% is higher than their overall precision.

A manual inspection of user corrections shows that users frequently did not correct or insert missing adverbs or adjectives whereas the general sentence structure was usually corrected.

5.2. Analysis

Overall, the relative improvement of transcript quality is much less than described by Munteanu et al. [5]. However, due to the lower initial error rates, the resulting WER is similar, supporting the observation, that a WER of 25% is somewhat acceptable for user and better, more fine-grained, corrections are perceived as too cumbersome.

Despite different degrees of familiarity with the lecture topics, all users performed quite similar. However, the precision of user edits is relatively low, much less than the 97% described by Yu [7].

Some loss in precision could be attributed to compounds and inflection in the German language and a user preference of making the lecture transcripts readable rather than completely correct appears to be a reasonable explanation. This agrees with the observation, that case and (compound) spelling was rarely corrected.

Especially the precision of less than 50% if an alternate hypotheses from the confusion network is chosen suggests that users will accept suboptimal corrections if they can be selected quickly. Nevertheless, characteristic lecture data errors were corrected manually if essential for the meaning. Phrases central to a lecture were continuously corrected even if they were mostly misrecognized.

6. Utilization for system adaptation

It is desirable that user correction do not only improve existing transcript, but rather improve future recognition performance. In this work, we investigated the use of corrected transcripts for system adaptation, compared with unsupervised adaptation on the CNC hypotheses.

Based on user corrections, a "consensus" transcript was created by using the most frequent user correction for each confusion network slot or the recognizer hypothesis if the slot has not been edited. Out-of-vocabulary words (OOVs) inserted by the users were split into existing sub-words in the vocabulary if possible. The rest was added to the vocabulary (without manual selection) with generated pronunciations.

Following the objective of improving simultaneous recognition of lectures, the "offline" correction should be used to adapt a single "online" system. Hence, the adaptation and evaluation is performed with a single MVDR system as opposed to the system combination of the first pass. Adaptation consists of vocal tract length normalization [18] and MLLR [19].

Evaluating the adapted system on unedited segments of the lectures shows that the low precision of user edits is problematic when using them as a basis for adapting models. When consensus transcripts are generated based on the user edits and used instead of the CNC output, small improvements on uncorrected data can be seen on content words, yet overall improvements in WER are not significant.

This lack of improvement compared to adaptation on the CNC output can be attributed to the relative sparsity of edited words and the heterogeneity of user edits, especially concerning compound treatment and typographical errors.

7. Conclusion

In this work, we presented a web interface for interactive correction of lecture transcripts and performed a user experiment to obtain information about quality and characteristics of user corrections without transcription guidelines.

User corrections improved the comprehensibility and quality of transcripts from a human perspective, i.e. for presentational purposes. This reduced the word error rate by a third to a level of 22%, which is, however, substantially worse than transcription quality. Especially the precision of user edits is relatively low at 77%, primarily due to errors in inflection, case and compound structure. This diminishes the usefulness of user-corrected segments for adaptation.

Future work will focus on utilization of corrections and refined adaptation methods. Additionally, it would be interesting to analyze user corrections on a larger scale in an actual setting and investigate the impact on subsequent machine translation.

8. Acknowledgements

'Research Group 3-01' received financial support by the 'Concept for the Future' of Karlsruhe Institute of Technology within the framework of the German Excellence Initiative.

The work leading to these results has received funding from the European Union under grant agreement no 287658.

9. References

[1] J. Glass, T. Hazen, S. Cyphers, I. Malioutov, D. Huynh, and R. Barzilay, "Recent progress in the MIT spoken lecture processing project," in *INTERSPEECH-2007*, 2007, pp. 2553–2556.

[2] C. Munteanu, R. Baecker, G. Penn, E. Toms, and D. James, "The effect of speech recognition accuracy rates on the usefulness and usability of webcast archives," in *Proceedings of the SIGCHI conference on Human Factors in computing systems*, ser. CHI '06. New York, NY, USA: ACM, 2006, pp. 493–502.

[3] C. Fügen, A. Waibel, and M. Kolss, "Simultaneous translation of lectures and speeches," *Machine Translation*, vol. 21, pp. 209–252, 2007.

[4] C. Chelba, J. Silva, and A. Acero, "Soft indexing of speech content for search in spoken documents," *Computer Speech & Language*, vol. 21, no. 3, pp. 458–478, 2007.

[5] C. Munteanu, R. Baecker, and G. Penn, "Collaborative editing for improved usefulness and usability of transcript-enhanced webcasts," in *Proceedings of the twenty-sixth annual SIGCHI conference on Human factors in computing systems*, ser. CHI '08. New York, NY, USA: ACM, 2008, pp. 373–382.

[6] C. Munteanu, G. Penn, and X. Zhu, "Improving automatic speech recognition for lectures through transformation-based rules learned from minimal data," in *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2*. Association for Computational Linguistics, 2009, pp. 764–772.

[7] G. T. Yu, "Efficient error correction for speech systems using constrained re-recognition," Master's thesis, Massachusetts Institute of Technology, 2008.

[8] J. Ogata and M. Goto, "Speech repair: Quick error correction just by using selection operation for speech input interfaces," in *INTERSPEECH-2005*, 2005, pp. 133–136.

[9] J. Ogata, M. Goto, and K. Eto, "Automatic transcription for a web 2.0 service to search podcasts," in *INTERSPEECH-2007*, 2007, pp. 2617–2620.

[10] J. Ogata and M. Goto, "Podcastle: Collaborative training of acoustic models on the basis of wisdom of crowds for podcast transcription," in *INTERSPEECH-2009*, 2009, pp. 1491–1494.

[11] M. Marge, S. Banerjee, and A. Rudnicky, "Using the amazon mechanical turk for transcription of spoken language," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*, march 2010, pp. 5270 – 5273.

[12] C. Lee and J. R. Glass, "A transcription task for crowdsourcing with automatic quality control," in *INTERSPEECH*, 2011, pp. 3041–3044.

[13] L. Mangu, E. Brill, and A. Stolcke, "Finding consensus in speech recognition: word error minimization and other applications of confusion networks," *Computer Speech & Language*, vol. 14, no. 4, pp. 373–400, 2000.

[14] S. Stüker, F. Kraft, C. Mohr, T. Herrmann, E. Cho, and A. Waibel, "The KIT lecture corpus for speech translation," in *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 2012, to appear.

[15] H. Soltau, F. Metze, C. Fügen, and A. Waibel, "A one-pass decoder based on polymorphic linguistic context assignment," in *Automatic Speech Recognition and Understanding, 2001. ASRU '01. IEEE Workshop on*, 2001, pp. 214–217.

[16] S. Stüker, K. Kilgour, and F. Kraft, "Quaero 2010 speech-to-text evaluation systems," in *High Performance Computing in Science and Engineering '11*, W. E. Nagel, D. B. Kröner, and M. M. Resch, Eds. Springer Berlin Heidelberg, 2012, pp. 607–618.

[17] K. Kilgour, C. Saam, C. Mohr, S. Stüker, and A. Waibel, "The 2011 KIT Quaero speech-to-text system for spanish," in *IWSLT-2011*, 2011, pp. 199–205.

[18] P. Zhan and M. Westphal, "Speaker normalization based on frequency warping," in *Acoustics, Speech, and Signal Processing, 1997. ICASSP-97., 1997 IEEE International Conference on*, vol. 2, apr 1997, pp. 1039 –1042 vol.2.

[19] C. Leggetter and P. Woodland, "Maximum likelihood linear regression for speaker adaptation of continuous density hidden markov models," *Computer Speech & Language*, vol. 9, no. 2, pp. 171 – 185, 1995.