Optimizing Computer-Assisted Transcription Quality with Iterative User Interfaces

Matthias Sperber[1], **Graham Neubig**[2], **Satoshi Nakamura**[2], **Alex Waibel**[1]
[1]Karlsruhe Institute of Technology, Institute of Anthropomatics and Robotics, Germany
[2]Nara Institute of Science and Technology, Augmented Human Communications Laboratory, Japan

Abstract

Computer-assisted transcription promises high-quality speech transcription at reduced costs. This is achieved by limiting human effort to transcribing parts for which automatic transcription quality is insufficient. Our goal is to improve the human transcription quality via appropriate user interface design. We focus on *iterative* interfaces that allow humans to solve tasks based on an initially given suggestion, in this case an automatic transcription. We conduct a user study that reveals considerable quality gains for three variations of iterative interfaces over a non-iterative from-scratch transcription interface. Our iterative interfaces included post-editing, confidence-enhanced post-editing, and a novel retyping interface. All three yielded similar quality on average, but we found that the proposed retyping interface was less sensitive to the difficulty of the segment, and superior when the automatic transcription of the segment contained relatively many errors. An analysis using mixed-effects models allows us to quantify these and other factors and draw conclusions over which interface design should be chosen in which circumstance.

Keywords: Computer-assisted transcription, user interfaces, quality-speed-tradeoff, mixed-effects models

1. Introduction

Manual speech transcription by human transcribers has risen in popularity in recent years (Ipeirotis, 2010). This may be surprising, because at the same time automatic speech recognition (ASR) technology has celebrated impressive advances. Even so, manual transcription through crowd-sourcing has been established as an attractive alternative and supplement to fully automatic transcription. In addition, there has also been a rise in computer-assisted transcription, in which automatic transcription is used as a starting point and human intervention is only needed when the ASR produces errors (Rodriguez et al., 2007). Both approaches promise better reliability than ASR, while being more affordable than fully manual expert transcription.

Our goal in this paper is to design a computer-assisted transcription user interface such that the outcome quality is optimized while avoiding unnecessary effort. The key interface feature we investigate is support for *iterative* transcription. This term is borrowed from iterative human computation processes (Little et al., 2010), in which humans solve tasks by improving upon a previously obtained solution. We consider computer-assisted transcription performed in an efficient segment-by-segment fashion, where only low-confidence segments are selected for manual transcription (Roy and Roy, 2009; Sperber et al., 2014b). Our iterative interfaces then provide the initial transcription as created by the ASR as a starting point for each segment, upon which the transcriber improves (cf. Figure 1). The benefit of the iterative interfaces is that the transcriber can simply use the initially correct parts from the ASR as-is, and focus attention on the problematic parts. Ideally, words that were recognized correctly by the ASR will not be changed, reducing the chance of correction errors. In addition, the iterative approach can assist transcription of parts that are difficult for the transcriber to understand by providing a first guess.

Computer-assisted transcription is traditionally performed by having the transcriber post-edit ASR results,

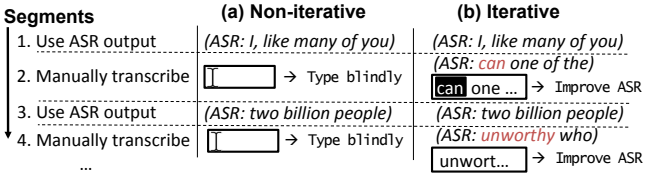

Figure 1: Computer-assisted transcription, low-confidence segments are manually transcribed with non-iterative (a) vs. iterative (b) interfaces. The actual utterance is "*I like many of you am one of the two billion people on earth who...*"

which is an iterative approach. However, a straightforward alternative would have been to type the correct transcription for each segment from-scratch. This would be the non-iterative approach. The difference of these two approaches is still poorly understood, so we conduct experiments to directly compare them in terms of quality and speed. Moreover, we examine two iterative extensions. First, the post-editing approach is enhanced by showing the ASR output with low-confidence words highlighted in red. This may help focus transcriber attention and decrease the chance of missing errors, a danger in traditional post-editing. Second, the from-scratch approach is extended into a retyping approach, again by displaying the confidence-highlighted ASR transcript above each segment. However, the text input box is initially empty and the transcriber is forced to retype every single word. Retyping also has the potential to focus transcriber attention and prevent missing errors.

The main contributions of this paper are introduction of the retyping interface approach, and an analysis of how retyping and the described three traditional iterative and non-iterative interfaces compare in terms of transcription quality and speed, in which we analyze how segment difficulty, mechanical effort, and other factors influence quality and speed. Specifically, we carry out a user study in which expert and non-expert participants transcribe automatically selected segments based on an ASR transcription, using the

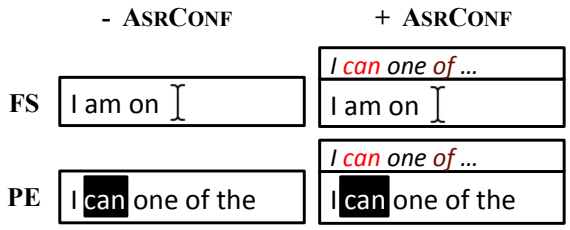

Figure 2: The four interface combinations: Typing from-scratch in an empty box (top), post-editing the ASR output (bottom), and unrolling the ASR output in sync with the audio while coloring low-confidence words in red (right).

four transcription interface designs. Our results show that non-iterative from-scratch transcription is a poor choice in general, even for skilled transcribers. All three iterative interfaces were roughly on par on average, but our retyping approach gave the best results for segments with high word error rate (WER) in the automatic transcript. This finding was consistent across transcribers. Savings in transcription time can be gained by switching to post-editing for segments with low ASR-WER.

2. Investigated Transcription Interfaces

The first interface feature we evaluate is typing from-scratch (FS) vs. post-editing (PE) the ASR output. Generally, post-editing needs less typing but requires additional effort for verification and navigation within the ASR output. Post-editing also carries the risk that the transcriber may miss some errors due to lack of attention.

The second feature (ASRCONF) displays the ASR output in a text label above the input field and unrolls it in sync with audio playback. ASR confidence scores are visualized by coloring words in shades between black (confident) and red (uncertain). Used for from-scratch transcription (FS$^+$), we obtain a novel retyping extension that offers the benefits of iterative transcription. Compared to post-editing, retyping may increase transcriber attention because the transcriber explicitly needs to (re-)type every single word. Using ASRCONF for post-editing (PE$^+$), the confidence visualization might prevent missing errors. FS$^-$ and PE$^-$ denote the traditional interfaces without ASRCONF.

The four possible combinations (FS$^-$, PE$^-$, FS$^+$, PE$^+$) are shown in Figure 2.

3. Experiment Design

We conducted experiments in order to find out which interface is best under what circumstances, regarding quality of the outcome and cost. We asked nine participants from different backgrounds to transcribe a selection of segments with varying length and difficulty. Transcription interfaces were alternated between segments. Segments were selected from automatically transcribed TED[1] talks. TED talks are short English talks directed to a general audience, presented by well-prepared speakers. The high quality of both audio and available reference transcripts allow for a well-controlled study with limited outliers and noise issues. The four talks in our evaluation set totaled 6,813 words or

[1]www.ted.com

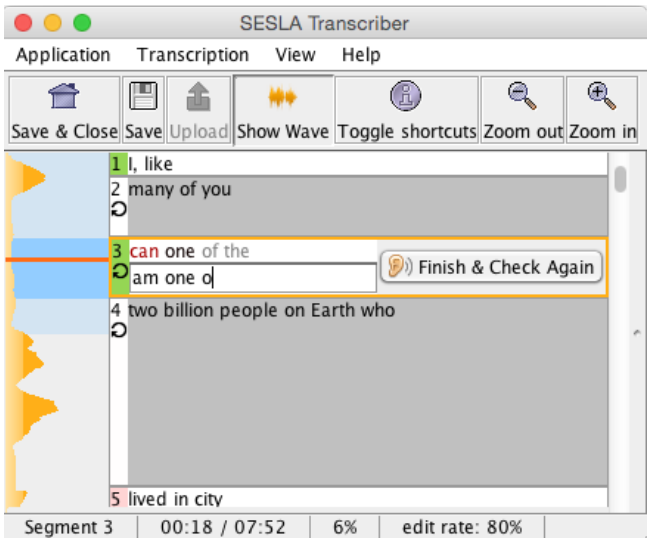

Figure 3: Screenshot of the transcription tool.

48 minutes of audio. They were selected from the IWSLT *dev2012* development set (Cettolo et al., 2014), for which high-quality reference transcriptions are available.

3.1. Transcription Data

As a starting point, we automatically transcribed these talks using a TED-optimized ASR system, described in detail in (Kilgour et al., 2014). This system showed state-of-the-art results in the 2014 IWSLT evaluation (Cettolo et al., 2014), which allows assessing if and by how much our computer-assisted transcription approach improves over a strong, fully automatic setup.

We focused on low-confidence segments for human transcription. This choice improves the balance in number of ASR errors between segments and makes our statistical analysis more robust. Moreover, it simulates a practical use-case in which no effort is wasted on transcribing segments with high ASR confidence. We selected these low-confidence segments using SESLA (Sperber et al., 2014b), which divides a transcript into segments that are of comfortable length for transcription, and consist mostly of low-confidence words. This method requires word-wise confidence scores, which we extracted from the lattice. We configured SESLA's optimization criterion such that 90% of the uncertainty mass would be transcribed. This resulted in 461 segments for transcription, with length between 1 and 15 words. These segments contained 53% of all words. The WER of the four talks averaged 11.3%. The WER for the high-confidence segments that were skipped by the transcribers was 2.5%, while for low-confidence segments that were chosen for transcription the WER was 21.2%.

3.2. Software

We used the tool SESLA Transcriber (Sperber et al., 2014a) to conduct our study (Figure 3). This tool takes a pre-segmented automatic transcription and displays the segments as boxes, aligned next to a vertically oriented waveform. Clicking a segment plays back the corresponding audio and allows typing the corresponding transcription. To convey enough context, segments chosen not to

be corrected are also displayed with their ASR output. The tool highlights misspelled words in red while typing, and suggests spelling corrections via a context menu.

3.3. Participants

We selected a heterogenous group of people as participants. In order of decreasing expertise, participants included skilled German transcribers with linguistic background (group LING), German computer science students (group CS), both with excellent English skills, and several Filipino workers with no higher education but who are fluent in English (group CROWD). We believe that the motivation of all participants was reasonably high, and that differences in transcription quality are primarily caused by transcriber skill, and less by lacking motivation or cheating attempts as is sometimes an issue in crowd sourcing.

The variety in participants allowed us to examine whether people of different backgrounds require different interface designs. Moreover, observations that are consistent among the participants will likely generalize to other groups of non-experts. All participants were asked to transcribe the same selection of segments in the same order. For the first participant, we randomly assigned one of the four input methods described above to each segment. For following participants, we rotated the input method for each segment based on the previous participant's segments. In this way, each particular segment is transcribed according to all input methods in a balanced fashion, and by each transcriber at most once.

3.4. Collected Data

Most transcribers finished the transcription of all four TED talks, although a small number of talks were not transcribed by everyone. In total, we obtained 3304 observations (transcribed segments) from 9 people. These observations consist of a number of features: transcriber/speaker/segment identity, input method, edit distances between ASR, correction, and reference, and transcription time. We lowercased corrections and removed punctuation. We also normalized spelling variants via hand-written rules derived from manual inspection of all observed confusion pairs. These steps are important to ensure that we only count actual errors in our analysis. We performed a thorough outlier detection, removing 359 segments with detected transcriber idleness or with extreme feature values. Both user interface demo and data are available on our website.[2]

3.5. Mixed-Effects Model Analysis

Each segment in our experiment is transcribed once by various transcribers with various methods. The resulting observations are thus not directly comparable, because we do not know whether differences are caused by transcriber characteristics or by experimental settings. More generally, our experiments involve "random" factors that are difficult to control for, and that potentially have a significant influence on our observations. In fact, this is a common problem in user studies. Recently, linear mixed-effects models (short: mixed models) have become popular as a convenient way of dealing with such situations. For instance, mixed models have been used for error analysis in ASR (Goldwater et al., 2010) and machine translation (Federico et al., 2014), and for analysis of post-editing for translation (Green et al., 2013).

Mixed models are specified by the following components:

- *Response variable*: The central quantity for which we wish to determine how it is influenced by other measured covariates. In our experiments, this will be the post-correction error rate or the transcription time.
- *Fixed effects*: Numerical or categorical attributes that influence the response variable in a meaningful way. In this paper, we assume a linear relationship. In the case of categorical variables, the assumption is that the observations include all values out of a finite set. Because categorical variables have no ordering, non-binary categorical variables are split into several binary variables. Our fixed effects include the user interface, segment length, number of errors, and other factors.
- *Random effects*: Categorical factors that are hard to control for or hard to understand. Generally, the observations include only a limited sample of values out of a large set of possible values. In our case, random effects are the particular transcriber, talk, and segment. In the simplest case, for each random effect a random intercept is estimated by the model. Thus, e.g. for each transcriber a mean correction accuracy is estimated to explain some of the observed variance that could otherwise only be explained by a general error term. Random effects are modeled to obey a Gaussian distribution with the observed sample mean and variance to be estimated. For transcribers, we extend the random effect to also account for variations in the slope of each fixed effect. For example, the strength of influence of the user interface on outcome quality differs between transcribers, a fact that we thus explicitly model.
- *Error term*: The variance in observations that is not explained by the random effects is finally modeled by the general error term.

We restrict ourselves to *linear* mixed models, because adding more complexity to the models increases the risk of unstable fits and inspection did not reveal any strong non-linearities. We also experimented with polynomial mixed models, but did not observe any model improvements.

We tested the significance of all fixed effects in our models via likelihood ratio test (for $p \leq 0.05$). That is, we built a null-model with the effect in question removed, and examined whether this significantly reduced the model likelihood. We used *R* (R Core Team, 2014) and *lme4* (Bates et al., 2013) to perform our mixed model analysis.

4. Results: Transcription Quality

Sufficient quality of the resulting transcript is often the most important requirement. Insufficient quality might render all cost and labor useless. As a case in point, we ob-

[2] www.msperber.com/research/lrec-iterative-gui

Method	ASR	FS$^-$	PE$^-$	FS$^+$	PE$^+$
Avg. WER	17.6	21.2	13.1	14.4	13.4
Avg. PER	15.4	13.3	8.2	9.2	8.3

Figure 4: Average WER/PER for different input methods according to fitted model.

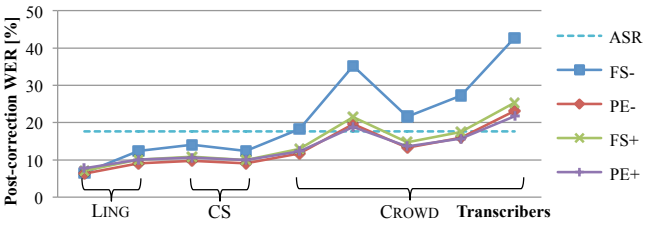

Figure 5: Estimated per-transcriber random effects of post-correction WER, for different input methods.

served that some of our transcribers actually *increased* the WER compared to the ASR for some interface setups.

To obtain a first overview over the results, we fitted a simple mixed model for the post-correction WER, with the transcription interface as a fixed effect, and random intercepts and slopes as described above. Figure 4 (WER row) shows the fitted intercepts for our fixed effects. It shows that using the ASR transcription as a starting point does in fact improve human transcription quality: The non-iterative, plain from-scratch (FS$^-$) interface was clearly outperformed by the three iterative interfaces, and was the only method that did not improve over the automatic baseline. PE$^-$ seems to be the best method on average, although the interface fixed-effect was not statistically significant between PE$^-$, FS$^+$, and PE$^+$ observations. Taking a closer look at the random fit for each transcriber (Figure 5) reveals that the poor performance of FS$^-$ was largely attributed to the non-expert group CROWD, who were not able to beat the ASR output with this method. However, note that even for several top transcribers, iterative interface design increased quality.

Next, we analyzed how presence or absence of ASR errors impacted the correction quality. We extended the previous mixed model by a new fixed effect, the ASR-WER, which interacts with the interface (i.e. the influence of ASR-WER is modeled to vary between interfaces). We also added this interaction as a random slope for transcriber identities. Figure 6 shows the fitted model. It can be seen that the presence of ASR errors considerably increased the chance of correction errors. The effect was strongly present for post-editing, indicating that transcribers missed more errors when more errors needed to be edited. It was present to a weaker degree even for plain from-scratch (FS$^-$), indicating that segments that were hard to transcribe for an ASR system were also hard for humans. Interestingly, FS$^+$ provided an appealing tradeoff between both: The impact of ASR errors was weaker, and the resulting quality better for segments with a WER of 29% or higher.

Given the variety in participants, is the sensitivity to the difficulty of a segment consistent across different transcribers? To answer this, we inspected the fitted random effects across transcribers. Figure 7 shows all transcribers' slopes for how fast post-correction WER grows when ASR-WER grows, concentrating only on FS$^+$ and PE$^+$. The sensitivity to difficulty varies considerably across transcribers, but there is no clear trend between the transcriber groups. It can be seen that sensitivity was consistently smaller for FS$^+$ than PE$^+$, regardless of transcriber skill. In fact, the slope of the ASR-WER was consistently smaller for FS$^+$ than for both PE$^{-/+}$, and larger than FS$^-$ for seven out of nine transcribers. Computing the crossing points, that is, the ASR-WER above which FS$^+$ yielded the better correction, gave a median of 19.5%, considerably lower than the sample mean of 29%. Given the linear nature of our model, the crossing points should not be taken as precise numbers, but more as a rule of thumb. In conclusion, retyping can be expected to be the superior interface for ASR-WERs above 20–30%. Given that such segments are usually more crucial to correct than the lower WER segments, it would also be a reasonable default choice of transcription interface.

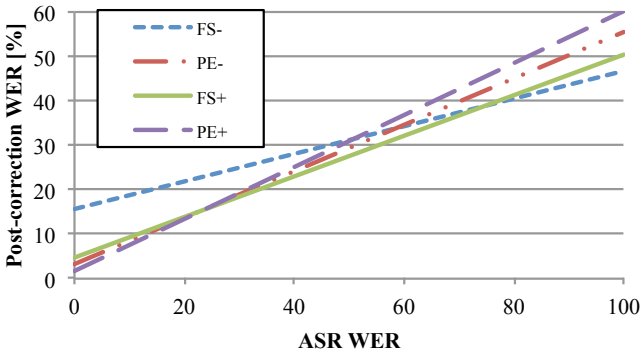

Figure 6: Influence of ASR WER on post-correction WER.

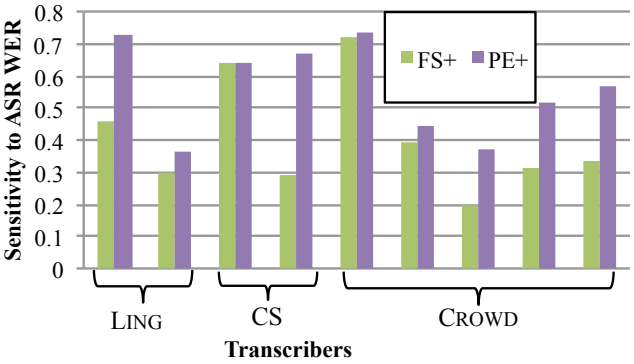

Figure 7: Transcribers' sensitivity to transcription difficulty for FS$^+$ and PE$^+$ input methods.

A question remains: Given the observed human transcription errors, how severe are these errors actually? Figure 4 can give us a clue by comparing the WER to the phoneme error rate (PER), a measure for acoustic similarity. These numbers were computed by replacing the segment WER by the segment PER in our first mixed model. It can be seen that according to PER, even FS$^-$ outperformed the ASR. The relative distances between the methods remain the same. In other words, humans appear better at choosing acoustically similar words than the ASR, even if the word is incorrect. Evaluating the semantic quality of

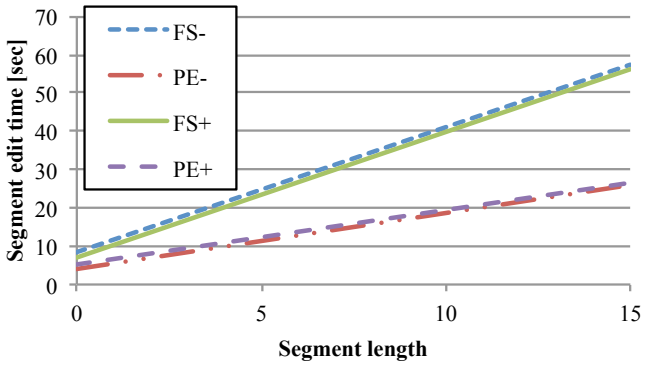

Figure 8: Linear mixed-model estimates illustrating how cost is affected by input interface and segment length.

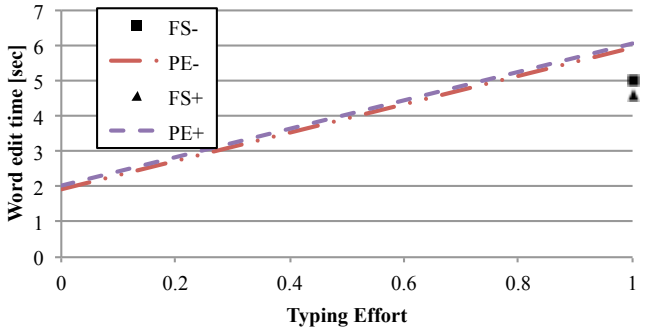

Figure 9: Linear mixed-model estimates illustrating how cost is affected by input interface, and by mechanical effort for post-edited input.

errors would also be interesting, but is outside the scope of this paper.

5. Results: Transcription Time

Besides quality, transcription speed is another important factor to consider. Faster interfaces can reduce cost and increase transcriber motivation. Therefore, we analyzed how transcription time is influenced by the transcription interface, while accounting for segment length. We estimated a mixed model with the segment transcription time as the response variable, and interface and segment length as interacting fixed effects. Random intercepts and slopes were as before. Figure 8 shows the fit of this linear model. It can be seen that while the ASRCONF feature had negligible effect on transcription speed, post-editing was consistently about twice as fast as from-scratch transcription. Closer analysis of the random intercepts and slopes for each transcriber showed that this observation held true for all transcribers.

According to the average segment WER, post-editing only required editing of 17.64% of the words in each segment on average. Given that from-scratch transcription required typing of all the words, its slower speed is not surprising. However, in Section 2. we hypothesized an inherent overhead for post-editing due to verification and navigation between errors that need correction. If true, segments with a high WER, for which the typing effort for post-editing is close to the from-scratch typing effort, might be faster to transcribe from-scratch. On the other hand, from-scratch transcription involves a higher amount of memorizing of the uttered words when ASRCONF is turned off, which might also cost time.

FS^-	PE^-	FS^+	PE^+	typing effort	error rate
0.91	1.81	0.56	1.91	3.90	0.98

Figure 10: Fitted model predicting *transcription time [seconds] per word*, comparing the effort inherent to user interface, mechanical effort, and influence of difficult words.

To validate this hypothesis, we built a more refined model that predicted the transcription time per word. The refined model used the input mode and an approximation of the mechanical typing effort as fixed effects. The mechanical typing effort is approximated as the proportion of words in the correction that have to be typed: in the from-scratch case this is all words, in the post-editing case we use the edit distance between correction and ASR. The effort is normalized by the reference segment length. The fitted model is displayed in Figure 9 and reveals that from-scratch editing is faster than post-editing for high PE typing efforts above 67%. In other words, for comparable mechanical typing effort from-scratch is faster, confirming our hypothesis about post-editing overhead. Moreover, FS^+ slightly but statistically significantly improves speed over FS^-, possibly because displaying the ASR hypothesis helps the transcriber recall the words uttered in the audio faster. However, bear in mind that most segments will have an edit rate much lower than 67%, meaning that the reduced typing effort of post-editing will usually outweigh its inherent overhead disadvantage.

Finally, we analyzed whether difficult words reduced transcription speed. We counted incorrectly transcribed words as difficult and other words as easy. The mixed model was as before, but with one additional fixed effect: the word error rate in the segment correction. All fixed effects were significant according to the likelihood ratio test. The fitted model is shown in Figure 10. We interpret these numbers as follows: After accounting for interface overheads, the mechanical effort for typing a word amounted to 3.9 seconds and difficult words slowed down correction by another 0.98 seconds. We conclude that the difficulty of a word has a strong impact on transcription time, a finding which might be useful for detection of human transcription errors.

6. Related Work

While from-scratch transcription and post-editing have been in use for many years, to the best of our knowledge this is the first systematic study of human transcription performance comparing these approaches as well as our extensions, or heterogenous user groups. However, several prior studies investigated related issues and established important prerequisites for our study.

6.1. Human Transcription Quality

Human transcribers are often categorized into professionals, i.e. trained transcribers, and crowd sourced workers. Professionals generally achieve the highest quality, but are expensive. Inter-transcriber disagreements have been

reported at 2–4% (NIST, 2009) and 5% (Novotney and Callison-burch, 2010). This disagreement results in part from ambiguities, and in part from a lack of domain knowledge that hinders transcription of specialized terms. High-quality transcripts are usually created in several passes, such that an initial expert transcription is verified and corrected by at least one other expert.

Crowd sourced workers on the other hand are inexpensive, but lack formal training. They may possess domain knowledge, such as students correcting lecture transcripts (Kolkhorst et al., 2012), or may lack domain knowledge, such as workers hired via Amazon Mechanical Turk (Novotney and Callison-burch, 2010). The mentioned works report WERs around 22% and 17% compared to reference transcripts. In the case of student transcriptions, a focus on correction of specialized terms that are most critical for understandability is reported (Kolkhorst et al., 2012). Multiple redundant crowd transcripts can be combined to yield near-expert quality (Audhkhasi et al., 2011), although larger improvements in ASR model training are reported by spending money on transcribing more data once, as compared to transcribing less data redundantly (Novotney and Callison-burch, 2010). Note that the WERs in our experiments cannot be directly compared, because we focus on especially difficult parts for transcription which are harder to transcribe for humans as indicated by our experiments and by Nanjo et al. (2006).

6.2. Transcription User Interface

Akita et al. (2009) conducted a user study with professional stenographers who post-edited ASR output. They report that post-editing time increases when the ASR produced many errors, and that the stenographers expressed a subjective preference towards typing from scratch when the WER of the ASR exceeded 25%. Nanjo et al. (2006) find that likewise typing from scratch takes more time for parts in which many ASR errors occur, indicating that parts that are difficult to transcribe for an ASR are difficult for a human as well. Our experiments confirm these trends and enable quantifying the various interacting factors. Bazillon et al. (2008) conduct an experiment including speech turn segmentation and transcription in a single step. They report that correcting an automatic segmentation and post-editing the transcription is much faster than segmenting and transcribing manually (from-scratch). Their experiments do not allow to draw conclusions as to how much the automatic segmentation and the automatic transcription contributed individually to this improvement. Valor Miró et al. (2015) make similar observations, and in addition investigate ways to exploit word-level confidence scores in a two-phase adaptation scenario.

Luz et al. (2008) propose a more advanced user interface for post-editing ASR output, in which low-confidence parts are highlighted, and alternative words are offered upon selecting an error. In a small user study, they report slightly higher accuracy at slightly increased cost for their user interface, compared to a plain text editor interface. Accuracy improvements are due to a smaller number of missed errors, possibly thanks to the confidence highlighting. Increased cost is caused by the more complex user interface. In our work, we have evaluated the confidence highlighting but omitted the complex user interface. We think that alternative lists are especially appealing for touchscreen devices, on which typing is cumbersome (Liang et al., 2014). Efficiency improvements over post-editing are reported for alternative lists in Japanese (Ogata and Goto, 2005), while for English, it is reported that only a third of all errors can be retrieved via alternative lists (Harwath et al., 2014). Moreover, Kolkhorst et al. (2012) report that alternative lists lead users to select suboptimal choices when the correct choice is not present in the list. Because of these issues, we refrained from including such more complex interfaces into our study.

Another popular method for error correction is respeaking (Vertanen and Kristensson, 2009), in which the utterance is repeated by a respeaker in a quiet environment and with a speaker-adapted ASR. Moore et al. (2004) point out that, while speech has a higher input rate, content creation rate (i.e., accounting for error correction) is much lower than typing on a keyboard. Sperber et al. (2013) show that respeaking by non-experts can be an option when ASR transcripts are to be improved, but not made perfect. However, respeaking requires recording equipment, a quiet environment, and a clear speaker. These factors that can often not be guaranteed for, especially in crowd-sourcing situations.

7. Conclusion

This work investigates the benefit of iterative user interface designs in the context of computer-assisted transcription with automatically chosen segments. A user study showed that a non-iterative, plain from-scratch transcription interface is clearly outperformed by our three evaluated iterative interfaces, regarding both quality and transcription time. This finding was consistent over transcribers of different skill. These three interfaces include traditional post-editing (PE^{-}), confidence-enhanced post-editing (PE+), and a novel retyping approach (FS^{+}). PE^{-} and PE^{+} behaved very similarly, indicating that the visualization of confidences has little effect on transcription. Quality was similar on average, but $\text{PE}^{-/+}$ yielded better quality for segments with low ASR-WER, and FS^{+} better quality for ASR-WER above 20–30%. $\text{PE}^{-/+}$ was considerably faster except for high edit rates above 0.67. Transcription quality of less skilled transcribers was especially sensitive to interface design.

In conclusion, FS^{+} is an appealing choice to ensure removal of segments with critically high WER, while the appealing quality of $\text{PE}^{-/+}$ is its cost-quality tradeoff. For an optimal quality-cost tradeoff, we suggest using post-editing for segments with low ASR-WER, and switching to retyping for higher ASR-WER. How to automatically predict which of the two methods should be used remains as future work.

8. References

Akita, Y., Mimura, M., and Kawahara, T. (2009). Automatic Transcription System for Meetings of the Japanese National Congress. In *Interspeech*, pages 84–87.

Audhkhasi, K., Georgiou, P., and Narayanan, S. S. (2011). Accurate Transcription of Broadcast News Speech Using Multiple Noisy Transcribers and Unsupervised Reliability Metrics. In *International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pages 4980–4983.

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2013). lme4: Linear mixed-effects models using Eigen and S4. *R package version*, 1(4).

Bazillon, T., Bazillon, T., Estève, Y., Estève, Y., Luzzati, D., and Luzzati, D. (2008). Manual vs Assisted Transcription of Prepared and Spontaneous Speech. In *Language Resources and Evaluation (LREC)*, pages 1067–1071.

Cettolo, M., Niehues, J., Stüker, S., Bentivogli, L., and Federico, M. (2014). Report on the 11th IWSLT evaluation campaign, IWSLT 2014. In *International Workshop on Spoken Language Translation (IWSLT)*, pages 2–17.

Federico, M., Negri, M., Bentivogli, L., and Turchi, M. (2014). Assessing the Impact of Translation Errors on Machine Translation Quality with Mixed-effects Models. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1643–1653.

Goldwater, S., Jurafsky, D., and Manning, C. D. (2010). Which words are hard to recognize? Prosodic, lexical, and disfluency factors that increase speech recognition error rates. *Speech Communication*, 52:181–200.

Green, S., Heer, J., and Manning, C. D. (2013). The efficacy of human post-editing for language translation. In *Conference on Human factors in computing systems (CHI)*, pages 439–448.

Harwath, D., Gruenstein, A., and Mcgraw, I. (2014). Choosing Useful Word Alternates for Automatic Speech Recognition Correction Interfaces. In *Interspeech*, pages 949–953.

Ipeirotis, P. G. (2010). Analyzing the Amazon Mechanical Turk marketplace. *XRDS: Crossroads, The ACM Magazine for Students*, 17(2):16.

Kilgour, K., Heck, M., Markus, M., Sperber, M., Stüker, S., and Waibel, A. (2014). The 2014 KIT IWSLT Speech-to-Text Systems for English , German and Italian. In *International Workshop on Spoken Language Translation (IWSLT)*, pages 73–79.

Kolkhorst, H., Kilgour, K., Stüker, S., and Waibel, A. (2012). Evaluation of Interactive User Corrections for Lecture Transcription. In *International Workshop on Spoken Language Translation (IWSLT)*, pages 217–221.

Liang, Y., Iwano, K., and Shinoda, K. (2014). An Efficient Error Correction Interface for Speech Recognition on Mobile Touchscreen Devices. In *Workshop on Spoken Language Technology (SLT)*, pages 454–459.

Little, G., Chilton, L. B., Goldman, M., and Miller, R. C. (2010). Exploring iterative and parallel human computation processes. In *ACM SIGKDD workshop on human computation*, pages 68–76.

Luz, S., Masoodian, M., Rogers, B., and Deering, C. (2008). Interface design strategies for computer-assisted speech transcription. In *Australasian Conference on Computer-Human Interaction Designing for Habitus and Habitat (OZCHI)*, page 203.

Moore, R. K., Court, R., and Street, P. (2004). Modelling Data Entry Rates for ASR and Alternative Input Methods. In *Interspeech*, Lisbon, Portugal.

Nanjo, H., Akita, Y., and Kawahara, T. (2006). Computer Assisted Speech Transcription System for Efficient Speech Archive. In *Western Pacific Acoustics Conference (WESPAC)*.

NIST. (2009). RTE Project. Technical report, NIST.

Novotney, S. and Callison-burch, C. (2010). Cheap , Fast and Good Enough : Automatic Speech Recognition with Non-Expert Transcription. In *North American Chapter of the Association for Computational Linguistics - Human Language Technologies Conference (NAACL-HLT)*, pages 207–215.

Ogata, J. and Goto, M. (2005). Speech Repair: Quick Error Correction Just by Using Selection Operation for Speech Input Interfaces. In *Eurospeech*, pages 133–136.

R Core Team, (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rodriguez, L., Casacuberta, F., and Vidal, E. (2007). Computer Assisted Transcription of Speech. In *Pattern Recognition and Image Analysis*, pages 241–248.

Roy, B. C. and Roy, D. (2009). Fast transcription of unstructured audio recordings. In *Interspeech*.

Sperber, M., Neubig, G., Fügen, C., Nakamura, S., and Waibel, A. (2013). Efficient Speech Transcription Through Respeaking. In *Interspeech*, pages 1087–1091.

Sperber, M., Neubig, G., Nakamura, S., and Waibel, A. (2014a). SESLA Transcriber: A Speech Transcription Tool That Adapts To Your Skill And Time Budget. In *Spoken Language Technology Workshop (SLT)*.

Sperber, M., Simantzik, M., Neubig, G., Nakamura, S., and Waibel, A. (2014b). Segmentation for Efficient Supervised Language Annotation with an Explicit Cost-Utility Tradeoff. *Transactions of the Association for Computational Linguistics (TACL)*, 2(April):169–180.

Valor Miró, J. D., Silvestre-Cerdà, J. A., Civera, J., Turró, C., and Juan, A. (2015). Efficiency and usability study of innovative computer-aided transcription strategies for video lecture repositories. *Speech Communication*, 74(September):65–75.

Vertanen, K. and Kristensson, P. O. (2009). Automatic selection of recognition errors by respeaking the intended text. In *Workshop on Automatic Speech Recognition & Understanding (ASRU)*, pages 130–135, dec.