Writing with Speech: A Qualitative User Evaluation Study

Lili Szabó*, Balázs Tárján* and Péter Mihajlik*,**

*Department of Telecommunications and Media Informatics, Budapest University of Technology and Economics
**THINKTech Nonprofit Ltd.
Email: {lili,tarjanb,mihajlik}@tmit.bme.hu

Abstract—This paper explores qualitative aspects of evaluating automatic speech recognition (ASR) based dictation. We propose an evaluation method of ASR-based dictation from the point of view of the user. To gather data for the study, we conducted experiments with possible target users (n = 6), asking them to compose the same text using ASR-based dictation software for the Hungarian language, and to repeat the same task using three different ASR systems. An inherent part of this task is correcting errors made by the system.

We found that users who had less experience with ASR, are biased towards keyboard & mouse writing mode, the one that they were more familiar with. Moreover, they overestimate the duration of text production using speech (ASR-based dictation) as a writing mode, when compared to keyboard & mouse writing mode. As expected, higher number of errors in the ASR output lead to longer error correction time, but somewhat surprisingly, longer edit time is in negative correlation with the number of errors left in the corrected text. That is, even if the time taken for correcting ASR errors was longer, the quality of the corrected text was worse than the one of the system with low recognition errors.

I. INTRODUCTION

Speech recognition as a writing tool is used by professional writers, writers with learning disabilities on a daily basis, and an increasing number of novice users opt for using the new technology to improve their writing. One of the main characteristics of speech recognition is the reduction of cognitive demands during the text production process, and lowering the physical effort present in keyboard & mouse typing. This allows cognitive resources to be allocated for other writing sub-processes, such as planning the text-to-be-produced, and revising the text that has been already produced. An ASR-based dictation system can be regarded as an application that facilitates inter-cognitive representation-bridging communication [1] as it mixes written and spoken form of natural languages, and may extend human cognitive capacities by letting the user focus on text production alone, without the hassle of typing. It also invites the main benefits of computer writing, namely non-linearity in the writing process. However, as long as speech recognition is not advanced enough to be 100% accurate, post-editing of errors is an inevitable step to reach the desired output. It is important to understand how users deal with this task, and how they perceive it.

This paper investigates the influence of the error correction process on the qualitative evaluation of ASR-based dictation systems. The evaluation of qualitative aspects of speech recognition is difficult, since user studies are costly, and user self-reports are mainly unreliable, therefore, there are no universally accepted qualitative evaluation criteria for speech recognition. We focus on how users perceive the performance of an ASR system through the error correction process. By error correction we refer to the correction of misrecognitions that are made by the software, all existing words in the lexicon. The data for the subjective performance evaluation was collected from ASR-based dictation system users in a controlled experimental setup, through an imitation of the error correction process which means switching between speech and keyboard & mouse writing modes.

In Section II we describe existing work relevant to this study. Section III argues the motivation and goals of this study, and outlines the methodology used. Section IV contains a detailed description of our ASR-based dictation system. Section V then presents the complete setup of the experiments conducted for this study. Section VI discusses the results, and we conclude our work in Section VII.

II. RELATED WORK

It was shown in [2], that the cognitive load reduces, when professional translators use dictation, and there is a significant productivity gain, which is measured in translated words per minute. [3] discusses that given the immediate visual feedback on the screen, automatic speech recognition based (ASR-based) dictation introduces non-linearity, thus recursivity in text processing, and allows the user to produce and edit text using several writing modes, including continuous speech and keyboard typing. When compared to classical dictation mode, the main benefit of ASR-based dictation is that the text appearing in the visual field can serve as references.

It was only in the past ten years that non-quantitative evaluation of speech recognition software emerged. Cognitive processes of writers using ASR as a writing mode during error correction were immensely studied in [4]. They also argue that as opposed to classical dictation (using dictaphone) and keyboard & mouse typing mode, speech recognition based dictation does not force writers to adopt a specific writing style. On the other hand, error correction is known to be a writing sub-process that increases the cognitive load during writing. Parallel to improving the performance of the recognizer, many error correction strategies emerged, see [5] and [6]. Our method to improve the error correction described in Section III.
III. MOTIVATION AND METHOD

It has been reported in many studies, that a word error rate (WER) even less than 10% of a speech recognition system creates user dissatisfaction, and post-editing errors generates frustration [5]. People prefer typing in texts to correcting errors made by the recognizer. To improve user experience of an ASR-based dictation system, it is essential to reduce the error correction time. One obvious way is to reduce WER. We hypothesize that error correction time can significantly be reduced, if the use of text processing commands, such as new line, and the verbal form other punctuation marks, such as comma for the ‘,’ mark, is enabled. This ensures getting the transcription output of the recognizer already in a structured format. Better readability probably reduces the cognitive load while post-editing the text.

A controlled experimental setup is needed to get a better understanding about how users perceive the performance of an ASR-based dictation system. In order to obtain qualitative evaluation of ASR-based dictation we compare different designs with the same functionality. We compare freely accessible speech recognition tools for Hungarian; these are currently the Google WEB Speech API (from now referred to as Google1), the dictation engine available on many applications for Android smart phones, powered by Nuance (from now referred to as Nuance2), and the system that uses VOXerver decoder provided by Speechtex3, described in Section IV, that we will refer to as Speechtex. Although text processing commands such as new line are available for English and for other main languages in the Google and Nuance systems, they are not implemented for Hungarian. Therefore we evaluate the Google and Nuance systems disregarding the punctuation characters.

The following features of the systems are getting evaluated in this study:

1) accuracy of the system: word error rate (WER)
2) time taken for error correction and post-editing (edit time)
3) time taken to type in the text (type time)
4) success rate of completing the task of inputting the text to the computer - number of errors computed from the Levenshtein distance of the target text and the resulting text (success rate, number of errors)
5) perceived accuracy of the system
6) perceived severity of the errors made by the system
7) perceived reaction time / responsivity of the system
8) user estimation for the duration of dictation & edit input compared to keyboard & mouse input

These features we use as evaluation measures. We differentiate between performance measures (1–4), and qualitative measures (5–8) that address the participants perceptions about the performance of the systems.

A. Prediction

We expect a strong correlation between word error rate (WER) and edit time, with WER possibly explaining a big portion of the edit time variance. We also hypothesize that when enabling voice activated text processing commands, the Speechtex system performs better (lower WER and shorter duration of edit time), and is perceived to be performing better on the task.

It is also very important to note that the Speechtex system is specific to the legislation domain, and was built with special respect to enabling verbal inputting of text processing commands and punctuation marks, whereas the other tools are for a more generic use. It is by no means our goal to perform a comprehensive quantitative and qualitative evaluation and comparison of the three systems.

IV. SYSTEM DESCRIPTION

As mentioned in Section III, there is no freely available ASR system for Hungarian that enables voice activated punctuation and text processing commands. Using the decoder and client provided by our industrial partner Speechtex Ltd., we built the language model and acoustic model, and the framework for combining and optimizing weighted finite state transducers (WFST), so that text processing commands would be part of the Speechtex system.

A. Corpus Acquisition

The corpus we used to build the language model on is a translation memory; a collection of small text segments and their translations. The publicly available DGT-TM corpus [7] (EURAMIS) is a parallel corpus of 23 official EU languages, consisting of translation units (TU) that contain individual sentences or sentence-like fragments, the source language being mainly English, with the translations produced manually. We extracted all Hungarian segments from the TU’s, resulting in (2.6 M) sentence or sentence-like units. The original texts were written for legal purposes, therefore the domain of the corpus is mainly legislation. The corpus was divided into train and a held out part, that we used for development and evaluation purposes, see details in Table I. The token and type counts differ greatly in their pre- and post-normalization state due to the extensive normalization described in detail in Section IV-B. There was a drop in type counts, as all punctuation characters were stripped from the previous words; the token counts increased after normalization for the same reason.

B. Normalization

The most common way of handling non-lexical vocabulary items (times, dates, numbers, abbreviations) is verbal-domain language modeling, where all non-lexical vocabulary items are expanded to their verbal form during pre-processing the training data using text normalization rules. Since we decided to enable voice activation of the verbal form of punctuation characters and other symbols, verbal-domain language modeling

---

2http://www.speechtec.com
3http://www.speechtex.com
was the more appropriate approach to choose. This included the following steps: sentence segmentation and recomposition, decapitalization of non-proper names at sentence-initial position, expanding abbreviations, acronyms and numeric entities, converting punctuation characters into their written verbal form. A vast amount of other non-lexical occurrences also had to be detected and removed.

The reverse process, denormalization (deverbalization) and case restoration of the non-lexical entities takes place at the decoding phase, explained in Section IV-E.

C. Acoustic Model

Since there was no sufficient amount of data at our hand to train a standalone, task-specific acoustic model, a speaker independent model was adapted for this the experiments. The 6121 tied-states (13 Gaussians per state) of the initial model was trained on 50 hours of manually transcribed radio broadcast recordings using decision trees and embedded Baum-Welch re-estimation implemented in the HTK toolkit [8]. Then supervised, MLLR/MAP adaptation was performed to fit the phone distributions to the development set consisting of 10 minutes of read speech from 8 speakers.

Word-to-phoneme mapping of the lexicon was performed using simple grapheme-to-phoneme rules [9]. The verbalization of special symbols (comma, sentence final stop, etc.) and the pronunciation of some frequently used abbreviations and foreign words were given manually. For terms having more than one valid pronunciation an additional weight was also assigned to each realization.

D. Language Model

Based on the verbal expended training corpus a statistical 3-gram model was estimated with SRI Language Modeling toolkit (SRILM) [10] by using modified Kneser-Ney pruning method. Neither language model pruning, nor vocabulary cut-off was applied, thus all word types of the normalized training text (645 K) can be found in the vocabulary of the speech dictation system.

E. Decoding

The recognition models were integrated into a triphone level WFST in order to make the search space compact [11]. Similar to the dictation systems powered by Nuance and Google, the Speechtex solution is also based on client-server architecture. The dictation client (VOXclient) extracts standard, 39 dimensional MFCC features (including first and second derivatives and energy) from speech and forwards them to a WFST-based decoder (VOXserver [12]) where a one-pass decoding process is performed. The server returns the most probable word sequence to the client in every 250 ms.

The denormalization takes place in the client based on a mapping dictionary we created during training text normalization and lists all the verbalized tokens coupled with their original, written form.

V. EXPERIMENTAL SETUP

Participants were invited to participate in a forty-minute experiment consisting of two parts, followed by a qualitative evaluation questionnaire.

A. Material

The text-to-be-read for the experiment was taken from the evaluation part of the DGT-TM corpus (see Section IV-A). It consists of 7 consecutive sentences in the legislation domain. We imitate the setup of a user who wants to produce a legal-themed text using ASR-based dictation mode, and keyboard-mouse mode for post-edition.

B. Participants

6 participants were selected for the experiments. 4 of the participants had already used an ASR system prior to the experiment, 2 of them had never used it and 2 of them were expert users. Participant demographics: 3 female and 3 male speakers, within the age range of 22-38, recruited from the Budapest area, all native speakers of Hungarian, speaking the standard accent. All participants were recorded in our speech lab, using VOXclient on a PC laptop at a 16 kHz sampling rate with 16-bit quantization.

C. Experiment Part I - Reading in Two Modes

First, participants were provided with a printed copy of the text, and asked to get themselves familiarized with it. Then, they were asked to read the text two times, in two different styles: reading the text in normal style (meaning reading running speech); reading the text in extended style, where the verbal form of the punctuation marks are also read out loud, so that ‘,’ becomes comma. We will refer to the output of the first setting as Text 1, and to the output of the second setting as Text 2 in the later sections. After reading the text two times, participants were then given the chance to play back and listen the recorded audio with their voice, and inspect the outputting process of the 3 systems separately, as if they were talking to them online. This observation phase helped them to get a better impression of the performance of all 3 systems.

D. Experiment Part II - Editing

In the second half of the experiments the participants were asked to edit the text output of all 3 systems (in the same text editor), to reach the format of the original/target text. Since text processing commands are not steadily available for Hungarian in the Google and Nuance systems, Text 1 of these systems were edited; for the Speechtex system Text 2 was edited by all participants. The editing task was performed on that system output that yielded in higher WER score, that is, the editing task was performed on Text 1 for Google and Nuance, and on Text 2 for Speechtex. We measured the time taken for the editing task (task completion time or edit time), the number of errors in the edited text (Levenshtein distance between the text edited by the participant and the original text), and the task completion success rate (success rate) calculated as follows:

\[
\text{success rate} = \frac{\# \text{ characters in text} - \# \text{ errors}}{\# \text{ characters in text}} \times 100
\]
E. Post-Test Questionnaire

To conclude the experiment, participants were asked to fill out an evaluation questionnaire about how they perceived (the ASR-based) dictation and the post-editing error correction tasks. It consisted of both system specific (1-3) and system independent statements (4-6), with 5 response options in Likert-style (ranging from 1 to 5), where the higher scores indicate stronger agreement with the statement. The scores would then get normalized to 0-4, with the scores for some of the statements reverted (subtracted from 4). The statements are listed below (translated from Hungarian):

1) The system recognized my words correctly.
2) The system made severe errors while I was dictating
3) The reaction time / responsiveness of the system is fast.
4) If writing was part of my everyday work, I would dictate instead of typing.
5) ASR-based dictation is not my cup of tea.
6) If I had to type in the same text from scratch, it would have been shorter, than dictating and editing it.

After completing all experiment tasks, half of the participants were asked to type the same text from scratch, reading the text from a printed copy. This was to imitate a possible real world scenario, where users vouch for something based on their intuitions and not on empirical evidence, and we did not want the real experience to interfere with the response given to statement 6.

VI. RESULTS

The calculation of WER was case insensitive for both Text 1 and Text 2. (Text 1 is the output of the system in a normal reading mode, Text 2 is the output of the system when the verbal form of the punctuation marks was also read.) For Text 1, WER was calculated using a reference text that did not contain punctuation characters and new line, as they were not expected to appear in the output. Although the Nuance system probably has a feature to recover commas, and after longer pauses sentence final stops paired with sentence initial recapitalization, this happens in less than 10% of the cases, so we did not change the reference text. For Text 2 WER was calculated including punctuation marks and new line. Both WER based on Text 1 and WER based on Text 2 were calculated for all 3 systems, but most system comparisons are based on the WER on the text that the editing task was based on. In Table II WER and edit time results are shown, displaying individual values by system and by the 6 participant. The inter-participant and intra-system standard deviation is relatively high for all systems, part of this can be explained by the outlying high values of WER scores of one participant/speaker with the ID of /2. The speech of this participant is fast, happens on a very low energy level, contains a lot of unreleased stops in word final position, vowel reduction, contractions - kinds of features that were challenging to all 3 recognizers. The average Text 1 WER score for Google and Nuance was lower than for Text 2, due to the many unrecognized verbal forms of punctuation marks. This corroborates with our predictions, and verifies the decision of using Text 1 for the editing task for Google and Nuance. The possible reason why Speechtex WER scores are higher on Text 1 than Text 2 is that our verbal-domain language model also contained the verbal form of all punctuation marks, therefore, when these are not part of the utterance, it is a different kind of language input, than the language model was trained on.

Figure 1 represents the relationship of WER and edit time visually. The plot also contains the regression line, the correlation coefficient being 0.66, which indicates a relatively strong correlation between WER and edit time, as expected. The left side plot of Figure 2 shows that the number of errors left after the editing task is higher in average for Google (41), than for Nuance (16) and for Speechtex (9). The barplot on the right side shows that the average edit time was also higher for Google (488 s), than for Nuance (452 s) and for Speechtex (248 s). The average pre-edit errors by system are 713.0, 491.76 and 273.6 respectively.

Figure 3 shows a detailed visualization of edit time and success rate by system and participant. One would expect that longer editing time compensates for the higher amount of errors in the system output text, but it seems not to be the case. Lower numbers of error (i.e. lower WER) in the input text for the editing task yields not only to shorter edit time, but also to lower numbers of error in the output of the editing task, i.e. higher task completion success rate in average. The edit time and success rate are in negative correlation (r = -0.47, p < 0.05), as well as WER and success rate: r = -0.71, p < 0.05. As the systems with lower WER (Speechtex and Nuance) also produce better readable transcription output - Nuance automatically inserts newlines and starts a new sentence after longer pauses, and Speechtex enables voice activated text processing commands and the input of punctuation marks - it could also be a factor, however we do not have empirical proof for that.

Figure 4 shows how the duration of dictating plus editing relates to keyboard-mouse typing. It was only in 4 of the cases of all 18 (above the red horizontal line on the plot, indicating the mean typing time), that the average typing time was of a shorter duration than the duration of dictating and editing the text together. Somewhat surprisingly, the participants still reported an average score of 3.1 for the dictation & edit time vs. keyboard input (typing). 5 is the strongest vote for dictation & edit time perceived as faster than typing in the same text. What we found (summarized in Table III), that the only explanatory factor for this perceptual bias was prior experience with the use of ASR software: the more experience the participant had with ASR-based dictation, the more they were likely to estimate the duration of typing to be slower, than the duration of dictation and post-editing together. Could possibly happen, that the switch between the two writing modes (speech and keyboard & mouse) for the dictation and editing task is perceived longer.
Correlation of WER and Edit Time by System and Speaker

A: Success Rate and Edit Time
B: Number of Errors and Edit Time

Fig. 1. The plot visualizes the relationship between the 2 main performance measures: WER and edit time of (values are given in seconds). The plot also contains the regression line and the variance explained ($R^2 = 0.66$).

Fig. 2. A: mean number of errors left post-editing and B: mean edit time by system.

Fig. 3. Edit time and A: task completion success rate, computed from B: number of errors left in the final output.

Fig. 4. The plot contrasts the measured time taken (seconds) for dictating plus editing the texts, and the measured time taken to type in the text from scratch. The red horizontal line indicates the average typing time of 3 participants who were asked to complete the typing task.

TABLE III. EXPERIENCE WITH ASR AND ESTIMATED DURATION OF DICTATION& EDIT TIME VS. KEYBOARD INPUT

<table>
<thead>
<tr>
<th>dictation is ... than typing</th>
<th>no prior experience</th>
<th>has experience</th>
<th>expert</th>
</tr>
</thead>
<tbody>
<tr>
<td>much slower</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>slower</td>
<td>1</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>equal</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>faster</td>
<td></td>
<td></td>
<td>2</td>
</tr>
<tr>
<td>much faster</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

A. Statistical Analysis

As mentioned earlier in Section III, one of the purposes of this study was to explore the relationship between WER, edit time and the qualitative performance measures. For all subjective evaluation variables (perceived accuracy, perceived severity of system errors and system responsivity) (rank based) correlation was measured, to see how they interact with WER and edit time (see Figure 5 for details). We used Spearman’s rank based measure of association to measure correlation between WER and perceived severity of the mistakes made by the system: $\rho = 0.68$; and to measure the correlation between WER and perceived accuracy of system: $\rho = -0.79$. The correlation between edit time and perceived accuracy was $\rho = -0.79$, and there was no significant correlation between edit
time and severity of mistakes. The low perceived responsivity / reaction time of the Nuance system is due to the design decision, that only allows the recognition output to appear on the screen after a longer (around 0.4 sec) pause.

To check our hypothesis about the Speechtex system performing better on this specific task, we also ran non-parametric paired one-tailed Mann-Whitney-U test [13], to compare the perceived performance variables (perceived accuracy, perceived severity of mistakes, perceived responsivity / reaction time) that are five point Likert items. We found that the only significant difference was between Speechtex and Google with respect to perceived reaction time / responsivity (p = 0.02), and between Speechtex-Nuance (p = 0.02) and Google-Nuance (p < 0.01) with respect to perceived reaction time / responsivity. Similarly, non-parametric paired one-tailed Mann-Whitney-U tests were ran to make a pairwise comparison between the WER of the systems, and found that the WER of Speechtex was significantly lower than that of Google and Nuance (p < 0.01 and p < 0.05 respectively), and the WER of Nuance was significantly lower than the WER of Google (p < 0.05). Since edit time is normally distributed, we ran paired one-tailed t-tests, and found that the edit time was significantly higher on the outputs of Google and Nuance compared to Speechtex (p < 0.01), but this difference was not significant between Google and Nuance.

VII. Conclusion

The major focus of this study was the evaluation of ASR-based dictation systems, using measures other than solely WER and involving target users in the evaluation process. The average WER and edit time was significantly lower on the Speechtex system - the system that enabled text processing commands and was trained on an in-domain language model. We emphasize that the comparison of the 3 systems was only for the specific purposes of dictation based text composition on a specific domain. One of the expected outcomes is that the lower the WER of the system, the lower the edit time and number of edit distances are in the edited text. The results also show that editing a recognizer output with higher number of recognition errors and in an unstructured and hard-to-read format takes longer time, and is also left with higher number of errors after the editing process.

Another interesting finding of this study was that participants with less experience with ASR-based dictation tend to overestimate the duration of dictation and error correction compared to keyboard inputting. The possible implications of this finding are yet to be explored.

ACKNOWLEDGMENT

Special thanks are due to all participants for their voice and patience. This work is partially supported by the projects TAMOP-4.2.2.C-11/1/KONV-2012-0013 (FuturICT.hu), GOP-1.1.1-11-2012-0377 (WEBRA TIME SAVE), PIAC-13-1-2013-0234 (Patimedia), KMR-12-1-2012-0207 (DIANA), AAL-08-1-2011-0001 (PAELIFE).

REFERENCES