Evaluation of the voice to text transfer in augmented conditions

Matej Janíček, Karol Velíšek, Radovan Holubek, Roman Ružarovský
Slovak University of Technology in Bratislava, Faculty of Materials Science and Technology in Trnava
Institute of Production Technologies
Jána Bottu 2781/25, 917 24
{matej.janicek, karol.velisek, radova.holubek, roman.ruzarovsky}@stuba.sk

Abstract. *The aim of the research was evaluation of voice to text transfer in augmented conditions for the use in the verbal control of industrial robots. Research builds on previous research, evaluation of voice to text transfer in different conditions, where the main problem of the voice to text transfer has been defined. It was clear from the results that the main problem of voice to text transfer is ambient noise. In this research, a device for eliminating ambient noise was used – the limiter. Simulation was designed and implemented to eliminate the main problem of voice to text transmission, the same procedure of simulation were used. The results of previous research were compared with the results of current research. Based on the analyzes, the result of this research was determined.*

Keywords. voice to text transfer, verbal control, evaluation

1 Introduction

We use human speech in everyday life for communication. Voice control, however, gets ahead in other places of our lives. We have all used voice recognition technology in our daily lives, many times without even thinking about it: automated phone books and directories, voice dialing on our mobile phones, and integrated voice commands on smartphones are just a few examples. Doctors and lawyers use voice recognition every day to dictate notes and transcribe important information. New uses of this technology include military applications, navigation systems, in-car speech recognition, "intelligent" houses designed with voice command devices, and video games that allow a player to command his units using their voice only. Since the 1920s we have been exploring machine control by voice. With a lot of research into modern technology, human-machine collaboration is becoming a reality (Windmann & Haeb-Umbach, 2009). Human-machine collaboration is essential in industry. If man and machine work closely together, we can use this connection, namely the intelligence and flexibility of man and strength and the infinite repeatability of the machine (Gustavssona et al. 2017). However, today's very rare introduction of machine voice control in the industry is certainly a sign of the lack of perfection, reliability and safety of such a control system (Rogowski, 2012). Speech recognition has been studied since 1950, the latest developments in computer and telecommunications technology have improved speech recognition capabilities (Kohanski at al., 2002). There are many researches and studies in the literature to improve speech recognition and voice control, e.g. robotic cell voice control (Rogowski, 2013), prosthetic arm (Gundogdu et al., 2018), wheelchair (Qadri & Ahmed, 2009) or robot manipulator (Jayasekara et al., 2008). However, none of these researches addressed the reliability and evaluation of speech to text. That is why a research that focused on the reliability of speech-to-text transmission has been implemented. This research builds on the results of research into the reliability of voice to text transfer.

2 Objectives and methods

The aim of the research was to determine the reliability and security of voice transmission to text in augmented conditions. Security is the most important aspect of any job. The reliability and security of voice transmission to text has addressed previous research. It was clear from the results that the main problem of voice to text transfer is ambient noise. In this research, a device for eliminating ambient noise was used = the limiter. The limiter is a device used in recording studios and radio studios to suppress ambient noise. The limiter used in the simulations suppressed ambient noise at speech input, thus eliminating the main problem of voice to text transfer - ambient noise. A simple simulation of voice to text transmission was compiled.
Following is the procedure of simulation:
- the same text of 100 words was read by the same human voice (the same speaker) into a headset microphone

- the human voice was transformed to text using a personal computer equipped with a sound board, limiter, transform software and headset microphone
- the original text and the transformed text were compared by personal computer
- the voice to text transfer was evaluated, in other words, we compared the printed text and the text in the computer

Fig. 1 shows the procedure of simulation.

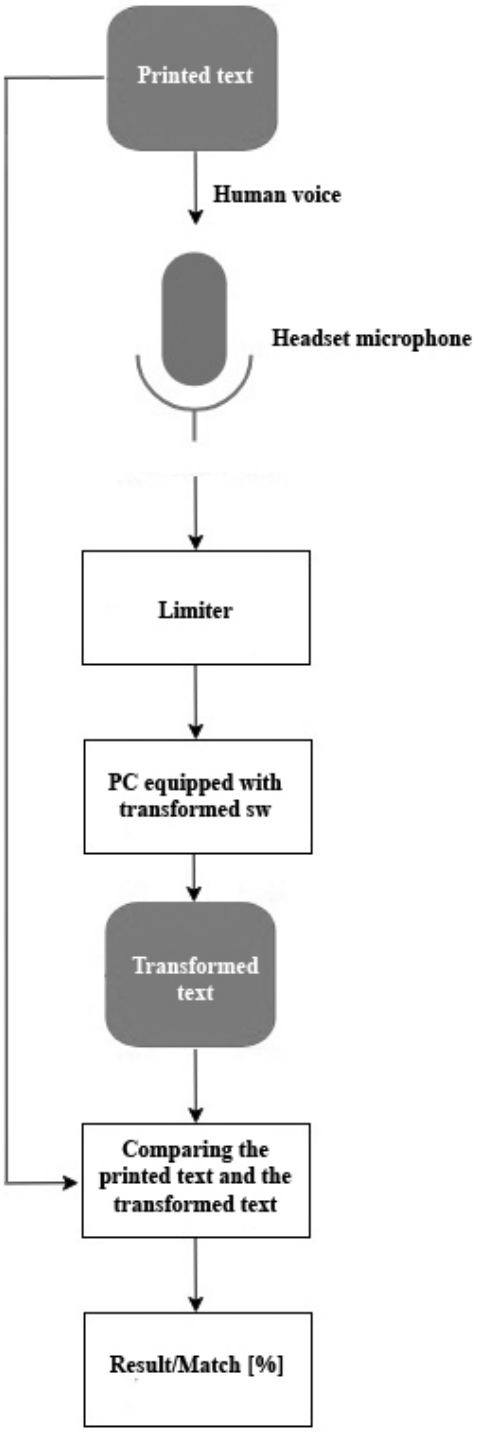

Figure 1 The procedure of simulation

Simulation conditions:
- three factors were used: X1 = speech speed
  X2 = speaker distance
  X3 = ambient noise
- each factor had two levels: Min/Max

Description of factors:
- factor X1: speaker was reading the same text at two different speech speeds
- factor X2: speaker was reading the same text at two different distances between the speaker´s mouth and a headset microphone
- factor X3: speaker was reading the same text under two different ambient noise conditions

Table 1. Simulation conditions

Factor	Mark	Min	Max
Speech speed [word/sec]	X1	1	2
Speaker distance [mm]	X2	50	100
Ambient noise [dB]	X3	40	90

Simulation implementation:

As already mentioned, each of the three factors affecting the voice to text transfer (X1, X2, X3) has two levels (Min, Max). In order to determine the impact of factors at different levels on speech to text, while identifying dependence of limiter on results of simulation, all possible combinations of factors had to be created at both levels. Factor combinations are shown in Table 2.

Table 2. Factor combinations

Factor			
X1	X2	X3	Mark of combination
Min	Min	Min	y1
Min	Min	Max	y2
Min	Max	Min	y3
Min	Max	Max	y4
Max	Min	Min	y5
Max	Min	Max	y6
Max	Max	Min	y7
Max	Max	Max	y8

In order to evaluate the results, 50 simulations of each combination (y1 to y8) of the three factors were performed. A total of 400 simulations were implemented (50 simulations of eight combinations =

400 simulations). The disadvantage of the simulation was the use of the human voice by only one speaker. During the voice to text transfer simulation, several free online software for the voice to text transfer (SpeechTexter, Speechnotes, Voice Notepad) was used. The software was randomly changed to avoid distortion of the simulation results if one software worked much better than the other, or if one software worked poorly, compared to others. Speech speed was measured with a simple timer, speaker distance between speaker's mouth and headset microphone using a caliper.

Ambient noise simulation was performed using a sound meter (simply application for android smartphone) and a conventional audio system. The audio system played a monotone sound, and the volume was adjusted to the desired value with the help of the sound level meter. When adjusting the volume, the sound meter was always as close as possible to the headset microphone. In this research, a device limiter was used. The device limiter was plugged in between headset microphone and personal computer, its job was to eliminate ambient noise at 300 mm or more from the headset microphone.

The same text of 100 words was chosen scientific text from a scientific article of course in English. Here is a short demonstration: Designing of conveyors and its control system with control program through the Virtual Commissioning design tool is very important in the digital era Industry 4.0. Virtual Commissioning was recently used to perform realistic virtual simulations in the early stages of development processes in automation of conveyors too.

The results of previous research, without using a personal computer equipped with a limiter, were compared with the results of current research. The simulation was carried out at the laboratory in Faculty of Materials Science and Technology in Trnava.

3 Results

After completing 50 simulations of the first combination of y1, the arithmetic mean of 50 simulations of the given the voice to text transfer match was calculated. This procedure was repeated for all combinations of y2 to y8. The results were recorded and evaluated.

Table 3. Average match of combinations of the voice to text transfer – previous research

Combination	X1	X2	X3	Average match
y3	1	100	40	90,60%
y7	2	100	40	90,52%
y5	2	50	40	89,98%
y2	1	50	90	89,62%
y8	2	100	90	77,06%
y1	1	50	40	75,30%
y6	2	50	90	73,72%
y4	1	100	90	72,98%

Table 4. Average match of combinations of the voice to text transfer – current research

Combination	X1	X2	X3	Average match
y3	1	100	40	92,26%
y7	2	100	40	92,18%
y5	2	50	40	92,08%
y2	1	50	90	92,04%
y8	2	100	90	91,96%
y1	1	50	40	91,88%
y6	2	50	90	90,82%
y4	1	100	90	90,62%

Table 3 and Table 4 shows the ordered combinations of factors, from the combination of the highest match to the least matched combination of the voice to text transfer. Table 3 shows results from previous research, table 4 from current research..

Subsequently, the influence of individual factors on the voice to text transfer was evaluated graphically.

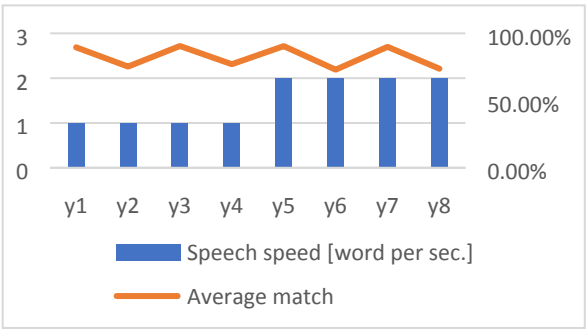

Figure 2 The impact of speech speed on the average match – previous research

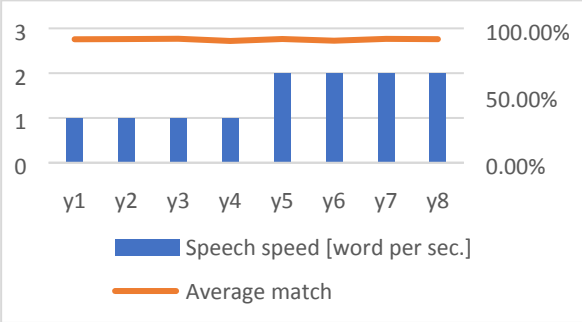

Figure 3 The impact of speech speed on the average match – current research

The impact of speech speed on the average match is shown in Fig. 2 and Fig. 3. In the combination of factors y1 to y4, speech speed 1 word per sec. was used in the simulation and in the combination of factors y5 to y8, speech speed 2 words per sec. was used. Fig. 2 shows results from previous research, Fig. 3 from current research

The impact of speaker distance on the average match is shown in Fig. 4 and Fig. 5. In the combination of factors y1, y2, y5 and y6, a speaker distance of 50 mm was used in the simulation, a speaker distance of 100 mm was used in the combination of factors y2, y3, y7 and y8.

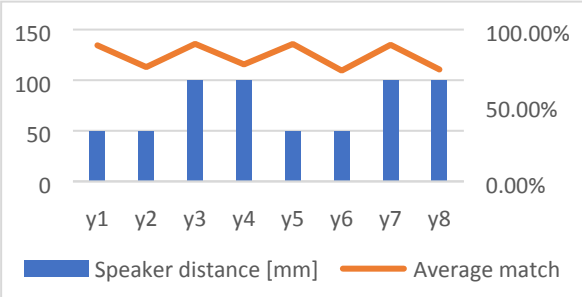

Figure 4 The impact of speaker distance on the average match – previous research

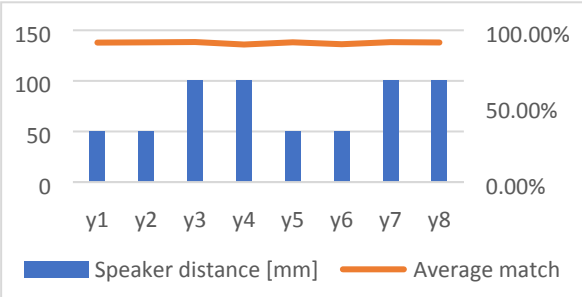

Figure 5 The impact of speaker distance on the average match – current research

The impact of ambient noise on the average match is shown in Fig. 6 and Fig. 7. In combination of the factors y1, y3, y5 and y7, ambient noise of 40 dB was used in the simulation, ambient noise of 90 dB was used in the simulation in combination of factors y2, y4, y6 and y8.

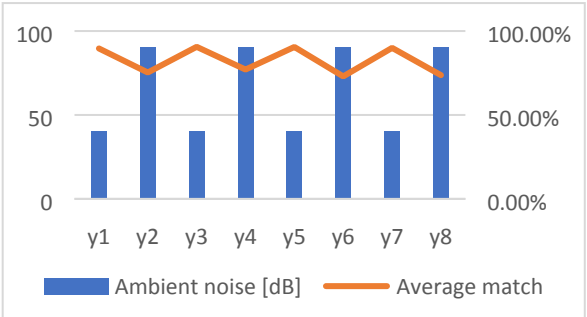

Figure 6 The impact of ambient noise on the average match – previous research

According to Fig. 6, ambient noise is clearly the most dependent on average match in previous research. The average match changed strategically whenever ambient noise changed. The main problem of the voice to text transfer has been defined and eliminated by the limiter used in this research to suppress ambient noise at speech input.

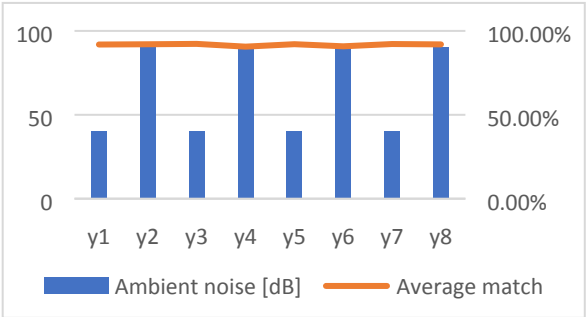

Figure 7 The impact of ambient noise on the average match – current research

Fig. 7 shows important difference between previous research and current research, when ambient noise was eliminated. The ambient noise changed strategically when average match not at all. The limiter directly helped improve speech to text transfer, but there still wasn't a 100% match between the printed text and the text in computer.

4 Summary

Current research has shown that limiter noise elimination has actually helped to voice to text transfer. In previous research were measured the lowest average match value of 72.98%. In current research were measured average match values from 90.62% to 92,26%. These are very acceptable values, but still not good enough to be used in practice.

5 Conclusion

The aim of this research was to determine the safety and reliability of voice to text transfer using an ambient noise eliminator. The human voice was transmitted by a microphone to a personal computer connected to a noise cancellation device and transformed into text. Subsequently, evaluation of voice to text transfer, comparison of original text and transformed text were made. The research was carried out in different conditions, at different speech speeds, speaker distances, ambient noise, and the results were compared with previous research. The simulation results in different conditions in the previous research have shown that ambient noise clearly has the greatest impact on the deterioration of voice to text. Every time the ambient noise changes, the voice to text transfer has changed strategically. Therefore, ambient noise was eliminated in the current research by device - the limiter used in the simulations suppressed ambient noise, and the results showed a significant improvement in voice to text transfer. The main benefit of this document is that noise elimination is the key to safer and more reliable voice to text transfer. However, what we have to remember is, that out of 400 performed simulations there was not a single one with a 100% match between the printed text and the text in computer.

Acknowledgments

This publication was created thanks to the support of the Ministry of Education, Science, Research and Sport of the Slovak Republic in the framework of the call for subsidy for the development project No. 002STU-2-1/2018 with the title „ STU as the Leader of the Digital Coalition.

References

Windmann, S., & Haeb-Umbach, R. (2009). Approaches to Iterative Speech Feature Enhancement and Recognition, *IEEE Transactions On Audio, Speech, And Language Processing*, Vol. 17, No. 5.

Gustavssona, P., Syberfeldta, A., Brewsterb, R. & Wangc, L. (2017). *The 50th CIRP Conference on Manufacturing Systems*, Procedia CIRP, 63, 396 – 401.

Rogowski, A. (2012). Industrially oriented voice control system, *Robot. Comput. Integr. Manuf.*, 28, 303–315.

Kohanski, M., Lipski, A., M., Tannir, J. & Yeung, T. (2002). Development of a Voice Recognition Program. Retrieved from *www.seas.upenn.edu/~belab/LabProjects/2001/be310s01t2.doc*

Rogowski, A. (2013). Web-based remote voice control of robotized cells. *Robot. Comput. Integr. Manuf.*, 29, 77–89.

Gundogdu, K., Bayrakdar, S. & Yucedag, I. (2018) *Journal of King Saud University – Computer and Information* Sciences, 30, 198–205.

Qadri, M. & Ahmed, S.A. (2009). *IEEE International Conference on Signal Acquisition and Processing*, 217– 220.

Jayasekara, B., Watanabe, K. & Izumi, K. (2008). *SICE Annual Conference*, 1, 2540–2544.