# Evaluations of the voice to text transfer in different conditions

*Matej* Janíček[1*], *Karol* Velíšek[1], *Radovan* Holubek[1], and *Roman* Ružarovský[1]

[1]Slovak University of Technology in Bratislava, Faculty of Materials Science and Technology in Trnava, Institute of Production Technologies, Jána Bottu 2781/25, 91724 Trnava, Slovak Republic

**Abstract.** The purpose of the research was evaluations of voice to text transfer in different conditons for the use in the verbal control of industrial robots. The research aims to find the best conditions to command an industrial robot using the human voice. A comprehensive study of existing problems and suggestions for simulation problems has been performed. Unlike some other works, it focused on the simply external problems to affect the voice to text transfer. The main problems of voice to text transfer (speech speed, speaker distance, ambient noise) has been established. The simulation using a personal computer equipped with a sound board and a headset microphone has been performed in all combinations of conditions. On the basis of the analyzes was establish the result of this research, the most suitable conditions and the worst conditions for voice to text transfer.

## 1 Introduction

Human speech is the most widely used means of communication between people. Using the human voice to control machines in the past seemed impossible. It also seemed impossible to control the equipment in the industry with a human voice. With a number of researches that date back to the 20s of the 20th century and the constant rapid advancement of modern technology, human-machine collaboration becomes a reality [1].

Human-machine collaboration is a very important aspect in the industry. If a person and a machine work closely together, we can take advantage of this connection, namely the intelligence and flexibility of man and strength and the endless repeatability of the machine [2]. However, today's very rare introduction of voice control of machines in industry is certainly a sign of the lack of perfection, reliability and safety of such a control system [3].

This research was conducted to determine the safety of voice to text transfer. The research was carried out under different conditions, the aim of the research was to find the most suitable and least suitable conditions for the voice to text transfer. The research results will be used in the future to test the voice control of an industrial robot.

In order for the voice to text transfer research to be realized, simulations were performed using the appropriate speech recognition device. Speech Recognition has been studied since 1950, the latest developments in computer and telecommunications technology have improved speech recognition capabilities [4].

* Corresponding author: janicek.matej6@gmail.com

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (http://creativecommons.org/licenses/by/4.0/).

There are many researches and studies in the literature aimed at improving speech recognition and voice control eg. voice control of robotized cell [5], prosthetic robot arm [6], wheelchair [7], or robot manipulator [8]. However, none of these researches has focused on the reliability and evaluation of speech to text.

## 2 Objectives and methods

The aim of the research was to determine the reliability and safety of the voice to text transfer. Safety is the most important aspect of any job. The same is true in the industry, where the risk of fatal injury is very high. If we want modern technologies such as voice control to be safe and reliable for humans, then what is the level of today's technology for human speech recognition and subsequent the voice to text transfer has to be explored. Various conditions/factors for the voice to text transfer have been selected and taken into account in the research, as it is complicated to manage equipment in the industry with human voice. A simple but still unrealized simulation of the voice to text transfer in selected conditions has been compiled for verification.

Description of the simulation:

- same text with 100 words has been spoken by same human voice (same speaker) into headset microphone
- human voice has been transformed to text using a personal computer equipped with a sound board, transform software and headset microphone
- original text and transformed text has been compared by personal computer
- the voice to text transfer has been evaluated

Simulation conditions:

- 3 factors has been used: X1 = speech speed
  X2 = speaker distance
  X3 = ambient noise
- Each factor had 2 levels: Min/Max

Description of factors:

- factor X1: speaker has been reading same text with 100 words by 2 different speech speed (1 word per sec./2 words per sec.)
- factor X2: speaker has been reading same text with 100 words by 2 different speaker distance between speaker´s mouth and headset microphone (50 mm/100 mm)
- factor X3: speaker has been reading same text with 100 words by 2 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 3 factors affecting the voice to text transfer (X1, X2, X3) has 2 levels (Min, Max). In order to determine the impact of factors at different levels on speech to text, while identifying the most suitable and least suitable speech-to-text conditions, 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 3 factors were performed. A total of 400 simulations were implemented (50 simulations of 8 combinations = 400 simulations). The disadvantage of the simulation was the use of 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.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 shows the ordered combinations of factors, from the combination of the highest match to the least matched combination of the voice to text transfer. Subsequently, the influence of individual factors on the voice to text transfer was evaluated graphically.

**Table 3.** Average match of combinations of the voice to text transfer

| Combination | Speech speed (word per sec.) | Speaker distance (mm) | Ambient noise (dB) | Average match |
|---|---|---|---|---|
| y3 | 1 | 100 | 40 | 90,60% |
| y5 | 2 | 50 | 40 | 90,52% |
| y7 | 2 | 100 | 40 | 89,98% |
| y1 | 1 | 50 | 40 | 89,62% |
| y4 | 1 | 100 | 90 | 77,06% |
| y2 | 1 | 50 | 90 | 75,30% |
| y8 | 2 | 100 | 90 | 73,72% |
| y6 | 2 | 50 | 90 | 72,98% |

The impact of speech speed on the average match is shown in figure 1. 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. According to this figure, speech speed has independent or minimal dependence on average match.

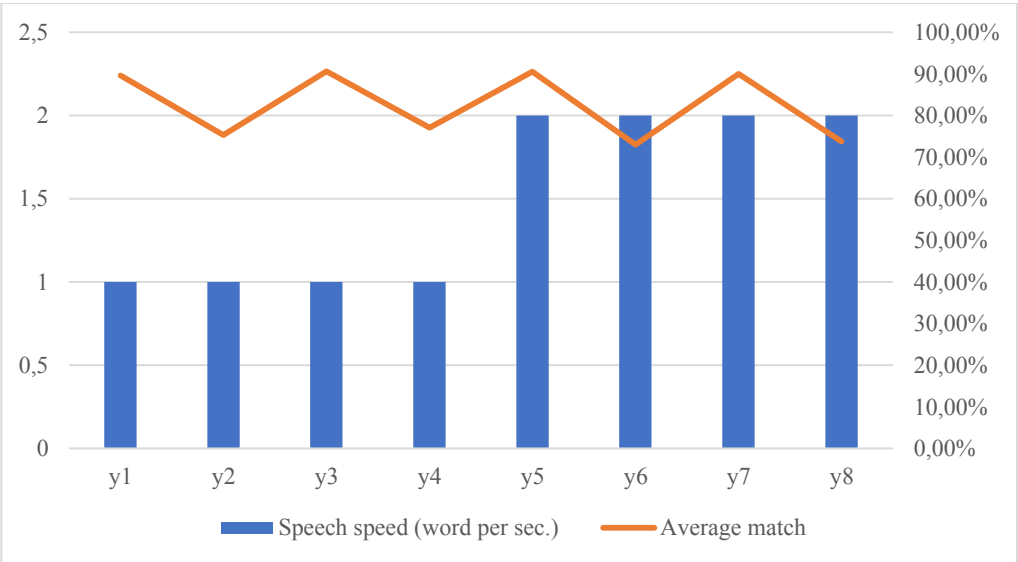


**Fig. 1.** The impact of speech speed on the average match

The impact of speaker distance on the average match is shown in figure 2. 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.

The average match has changed strategically in cases where speaker distance has not changed at all. According to this figure, the speaker distance has an independent or only minimally dependent effect on the average match.

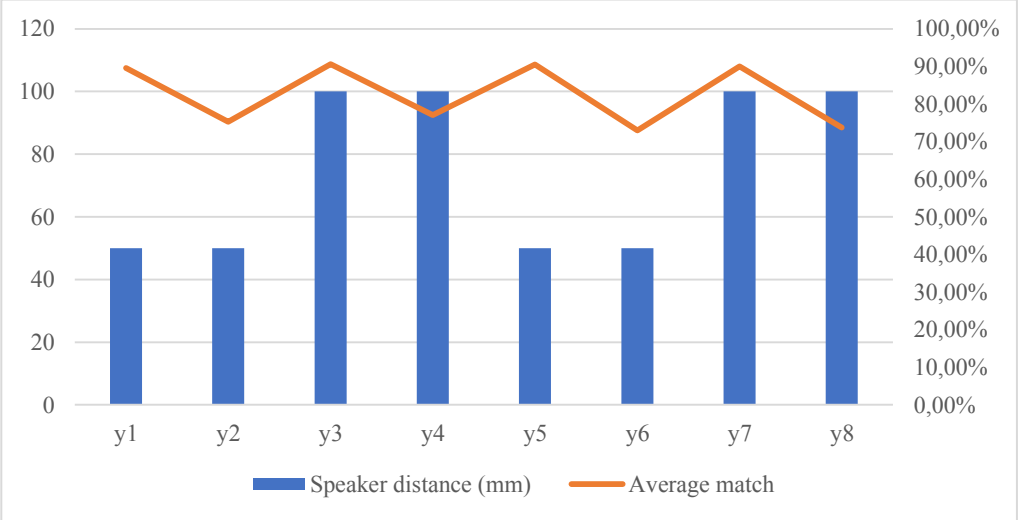


**Fig. 2.** The impact of speaker distance on the average match

The impact of ambient noise on the average match is shown in figure 3. 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. The average match changed strategically whenever ambient noise changed. According to this figure, ambient noise is clearly the most dependent on average match.

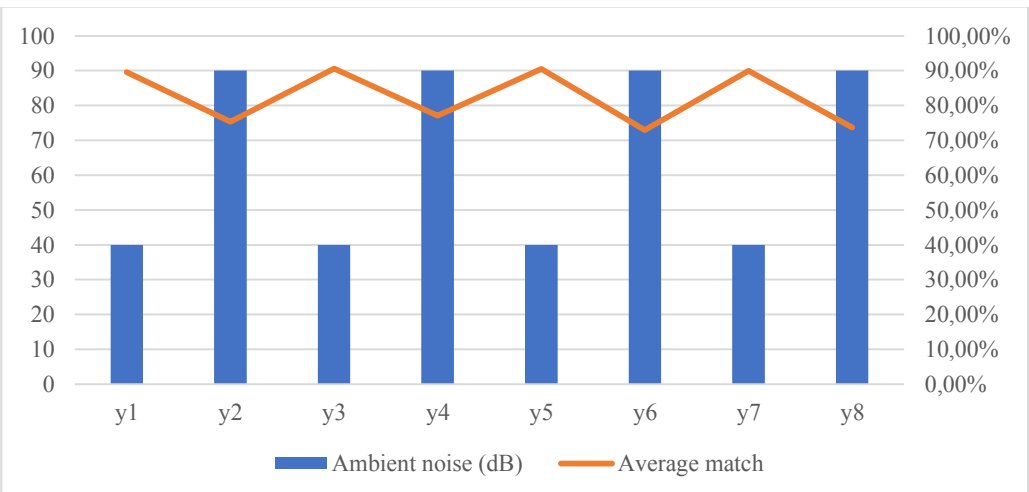


**Fig. 3.** The impact of ambient noise on the average match

## 4 Summary

Evaulations of the voice to text transfer:
- the most suitable conditions of the voice to text transfer: speech speed 1 word per sec., speaker distance 100 mm, ambient noise 40 dB, average match of the voice to text transfer under these conditions was 90.60% of transferred words.

- the worst suitable conditions of the voice to text transfer: speech speed 2 words per sec., speaker distance 50 mm, ambient noise 90 dB, average match of the voice to text transfer under these conditions was 72.98% of transferred words.

The results of the simulations showed the assumed strategic impact of ambient noise on the voice to text transfer.

## 5 Conclusion

The purpose of this research was to determine the safety and reliability of the voice to text transfer. Human voice was transmitted by a microphone headset to a personal computer equipped with a sound board and transformed to text. Subsequently, evaluations of voice to text transfer was performed, compare original text and transformed text. The research was carried out in different conditions, different speech speed, speaker distance, ambient noise to find the most suitable and least suitable conditions the voice to text transfer.

The results of the simulations in different conditions have shown that ambient noise clearly has the greatest impact on the deterioration of the voice to text transfer. Every time when ambient noise changed, the voice to text transfer changed strategically. Other 2 conditions, speech speed and speaker distance affect voice to text transfer only minimally.

The most suitable and least suitable conditions of the voice to text transfer have been identified. The main contribution of this document is that even under the most suitable conditions, the voice to text transfer has achieved an average match of 90.60% transformed words. This has proven the current unreliability hypothesis (less than 100%) and not the great safety of use and deployment of verbal control in the industry.

The future plans comprise the following:

- Develop a plan to eliminate ambient noise at voice control, that is common in any industry
- Utilize advanced software and hardware for simulation of the voice to text transfer

The research work reported here was made possible by the project KEGA-021STU-4/2018. Development of a laboratory for the design and maintenance of production sys-tems supported by the use of Virtual Reality

## References

1. S. Windmann, R. Haeb-Umbach, IEEE Transactions On Audio, Speech, And Language Processing, **17** (2009)
2. P. Gustavssona,, A. Syberfeldta, R. Brewsterb, L. Wangc, Procedia CIRP **63**, 303–315 (2017)
3. A. Rogowski, Comput. Integr. Manuf. **28** ,303–315 (2012)
4. M. Kohanski, A. M. Lipski, J. Tannir, T. Yeung, *Development of a Voice Recognition Program*, available at: http://www.seas.upenn.edu/courses/belab/LabProjects/2001/be310s0 1t2.doc (2002)
5. A. Rogowski, Comput. Integr. Manuf. **29**, 77–89 (2013)
6. K. Gundogdu, S. Bayrakdar, I. Yucedag, Comp. and Inf. Sc. **30,** 198–205 (2018)
7. M. Qadri, S.A. Ahmed, IEEE International Conference on Signal Acquisition and Processing, 217– 220 (2009)
8. B. Jayasekara, K. Watanabe, K. Izumi, SICE Annual Conference, **1**, 2540–2544 (2008)