Comparison of Speech Recognition and Natural Language Understanding Frameworks for Detection of Dangers with Smart Wearables*

Dariusz Mrozek[1][0000−0001−6764−6656], Szymon Kwaśnicki[1], Vaidy Sunderam[3][0000−0002−5128−7852], Bożena Małysiak-Mrozek[2][0000−0003−4977−4915], Krzysztof Tokarz[2], and Stanisław Kozielski[1]

[1] Department of Applied Informatics, Silesian University of Technology Akademicka 16, 44-100 Gliwice, Poland
dariusz.mrozek@polsl.pl
[2] Department of Graphics, Computer Vision and Digital Systems, Silesian University of Technology, Akademicka 16, 44-100 Gliwice, Poland
krzysztof.tokarz@polsl.pl
[3] Department of Computer Science, Emory University, Atlanta, GA 30322, USA
vss@emory.edu

Abstract. Wearable IoT devices that can register and transmit human voice can be invaluable in personal situations, such as summoning assistance in emergency healthcare situations. Such applications would benefit greatly from automated voice analysis to detect and classify voice signals. In this paper, we compare selected Speech Recognition (SR) and Natural Language Understanding (NLU) frameworks for Cloud-based detection of voice-based assistance calls. We experimentally test several services for speech-to-text transcription and intention recognition available on selected large Cloud platforms. Finally, we evaluate the influence of the manner of speaking and ambient noise on the quality of recognition of emergency calls. Our results show that many services can correctly translate voice to text and provide a correct interpretation of caller intent. Still, speech artifacts (tone, accent, diction), which can differ even for each individual in various situations, significantly influences the performance of speech recognition.

Keywords: Internet of Things · Cloud computing · Natural Language Processing · wearable sensors · intention recognition · speech recognition · older adults

* This work was supported by pro-quality grant for highly scored publications or issued patents (grant No 02/100/RGJ21/0009), the professorship grant (02/020/RGP19/0184) of the Rector of the Silesian University of Technology, Gliwice, Poland, and partially, by Statutory Research funds of Department of Applied Informatics, Silesian University of Technology, Gliwice, Poland (grant No BK-221/RAu7/2021).

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77970-2_36

1 Introduction

In many developed countries, the population is aging, due partly to longer life expectancy, which in turn is partly due to better medical care [1]. Among the elderly, human independence decreases, and the risk of disability increases. Despite family support and geriatric care, the elderly increasingly stay alone at home most of the time [27]. In the event of a sudden deterioration in health or an accident, they are often unable to call for help. In the event of an immediately life-threatening condition, e.g., myocardial infarction, stroke, or hypoglycemia, self-call for help is often impossible, and every minute of delay in implementing appropriate treatment carries irreversible consequences, including the patient's death. Smart wearable devices can help seniors (and others) in such situations by giving them a simple, technology-assisted, way to call their loved ones or request medical assistance.

The Internet of Things (IoT) has become an essential technology that allows people to monitor themselves during daily activities. In particular, there is growing interest in using IoT devices for personalized healthcare, monitoring older people and children, and detecting potential dangers in their lives [25, 29]. Smart wearable devices, like smart bands and smart bracelets, can gather information about vital health parameters, including pulse, ECG, body temperature, blood pressure, and oxygen saturation. Smartwatches and mobile phones can detect falls and notify caregivers automatically when fall is detected [24]. External sensors and cameras can monitor a person constantly and raise the alarm when the situation is dangerous. However, sometimes it is much easier for the monitored senior to call for help than to monitor all possible life parameters and vital signs with such wearable devices and external sensors. A smart solution for this purpose can be based on an IoT personal device (like a pendant, necklace, or smart band) that can register the voice of the monitored person, analyze it or transmit it for analysis after connecting to a data center (Fig. 1). Such technologies can be especially useful in (a) providing voluntary additional information to physiological measurements, and (b) distinguishing between false and true alarms.

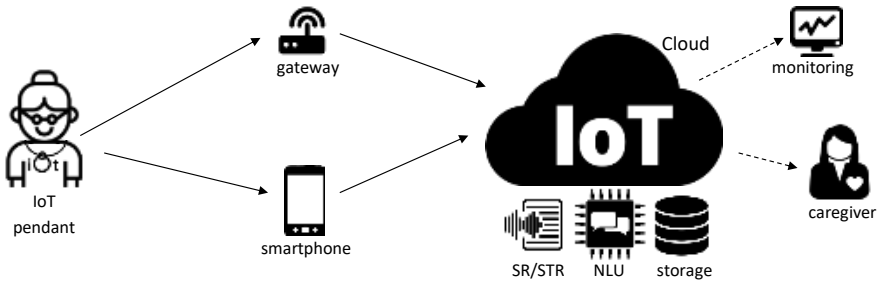

Fig. 1. Schematic diagram of monitoring an older adult with an IoT pendant registering voice and transmitting it for the analysis to a data center in the Cloud.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77970-2_36

In this paper, we analyze the capabilities of various speech recognition and natural language understanding frameworks in the automatic detection of situations when a monitored person is requesting help. We rely on the architecture of the experimental environment, where the detection of attention-requiring situations occurs in the central Cloud-based telemedicine system for monitoring and data processing.

2 Related Works

The most natural way of communication between humans is speech, which has led to a focus on spoken language as an important human-machine interface [8, 13]. Computer technology can understand phrases and whole sentences using spoken language understanding (SLU) systems. Such systems typically consist of a pipeline of two main modules – automatic speech recognition (ASR) followed by natural language understanding (NLU) [4].

Automatic speech recognition (ASR), also called speech to text recognition (STR), or speech to text analysis is an active area of research for many years. Currently available systems can recognize words in different noise conditions, spoken by different people in many languages, with high accuracy. Although their level of recognition does not achieve 100% they have been implemented in many solutions for everyday life usage, i.e., as the user interface to computers or smartphones, in smart homes [21], in telecommunication [18], and in the health care sector for automatic generation of medical documentation [11] or recording medical interview data [17]. STR systems give accessibility options for people with disabilities [18]. For example, Dimauro et al. presented research on using STR in measuring speech impairment in Parkinson's disease [11].

Natural Language Understanding (NLU) is one of the biggest challenges in computer science. Currently, recognizing separate words is quite simple but understanding their meaning while spoken in conjunction with other words is a complex task for computers [5]. To prevent misunderstandings, the Controlled Natural Language (CNL) approach can be used. Strong CNL has well-structured semantics, and words have a single meaning; it can be equivalent to formal language [26]. Weak CNL allows more than one word meaning but specifies some restrictions on the construction of sentences to avoid ambiguity [5]. It must be noted that in emergency situations, patients may not be able to utter a sentence according to the rules, placing CNL effectiveness in doubt. Although properly-recognized sentences can be highly informative, additional information can be obtained with emotion detection [9]. This can help when patients cannot speak a complete sentence, but utter emotionally short phrases or separate words.

NLU systems are implemented using deep neural networks (DNN) [10] or recurrent neural networks [12]. The continuing development of cloud and networking technology has driven the growth of a number of cloud-based services with artificial intelligence services, including such DNN-based services, and has allowed moving the voice recognition process to the Cloud. The first commercially available voice assistant (VA) was Apple's Siri launched in 2011 [3]. Sub-

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77970-2_36

sequently, competitors created their own systems, with the Microsoft Cortana (2014), Amazon Alexa (2014), and Google Assistant (2016) being the most popular. Cloud NLP is more effective than voice recognition technology built into end-devices [3] for several reasons: the possibility of installing custom applications; constant learning of new words and phrases from many users; and of course, computing power.

A comparison of the most popular VAs by Lopez et al. [19] shows that it is a non-trivial task to obtain both naturalness and correct operation at the same time. According to their research, the most natural but less accurate is Google Assistant, while Apple's Siri is the most accurate but the least natural [19]. Ammari et al. [2] surveyed everyday usage of voice assistant systems. Analyzing results, they found that the commands most often used with voice assistant systems are used for playing music, hands-free searching, and controlling IoT devices.

In the literature, there are many interesting examples of successful voice-controlled systems implementation. Many of them are created with open-source tools, including Node-RED, IFTTT, and well-known communication protocols like MQTT [16]. An example of smart home implementation based on Android is presented in [14]. In [3] Austerjost et al., a system to control real laboratory equipment with voice commands is presented. Using their system, it is possible to read measurements from sensors, report status, laboratory equipment parameters, and read operating procedures. In Mitrevski [23], the author describes events where the conversational interface used by a four-year-old child helped save his mother's life. The current research focuses mainly on methods of natural language processing (NLP) that allow recognizing not only simple commands but also conditional or personalized actions [20], and complex phrases with the ability to ask clarifying questions.

In our project, we adopt a different outlook that complements the descried approaches – we propose to build a system that acts as a personal assistant similar to home assistants, but focuses on the aspect of medical care and calling for help. The designed system has several functionalities, which are described in the next chapter.

3 Personal Assistant with Cloud Voice Analysis

A personal voice assistant is an IoT device that detects a dangerous event by registering the voice of a person that utters a call for help sentence or phrase. The assistant can be a band, pendant, or another type of device that an older person can easily carry. This mobile device directly or indirectly accesses the Internet to connect to a data center located in the Cloud. The main task of this device is to listen to the surrounding environment in standby mode. After detecting sounds from the environment, the sounds are analyzed to determine if they contain a speech recording or just irrelevant background noise. If the recording is classified as a voice, it is processed to highlight the desired features and remove unnecessary silence and surrounding sounds. Then the recording is

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77970-2_36

converted to a fixed file type with a certain quality (kbps) and uploaded to the cloud data center for further analysis, as shown in Fig. 2.

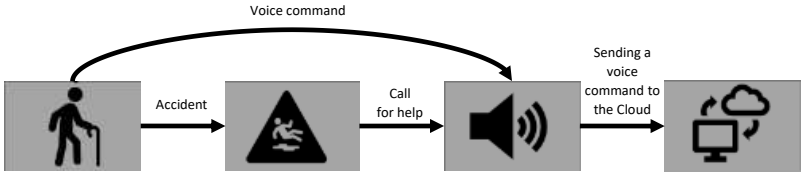

Fig. 2. Main operational phases in a call for help system with a data center for voice analysis located in the Cloud.

After receiving the speech recording by a dedicated application in the cloud data center, specific system modules begin intelligent processing of the voice command to determine if it contains a call for help. The intelligent processing of voice recordings consists of two steps (Fig. 3):

- the transcription of a voice command into text,
- appropriate analysis of the command represented in text, including intention recognition.

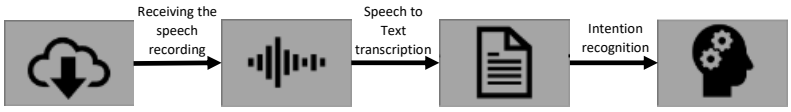

Fig. 3. Steps of voice command analysis.

Among the command recognition techniques, we use a combination of automatic speech recognition and information extraction using intent and entity (key elements in a sentence) recognition approaches. Among the systems found in the literature, we did not find any that use assistant mechanisms. Moreover, research has shown that older people have problems with commands that require strict construction [22]. Their statements differ from the standard ones by interwoven pauses and the addition of polite phrases (e.g., *Please find ...*). Therefore, the method used must have some flexibility.

The last step is the appropriate reaction to the recognized intention (Fig. 4). If the intention is to call for help, the caregivers mentioned in the command are notified by sending a specific message type. Moreover, in case of critical level messages indicating a life-threatening event, medical assistance can be called. Both

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77970-2_36

the positive (call for help cases) and negative results of the situation classification are recorded in the database for iterative training of the Natural Language Understanding models.

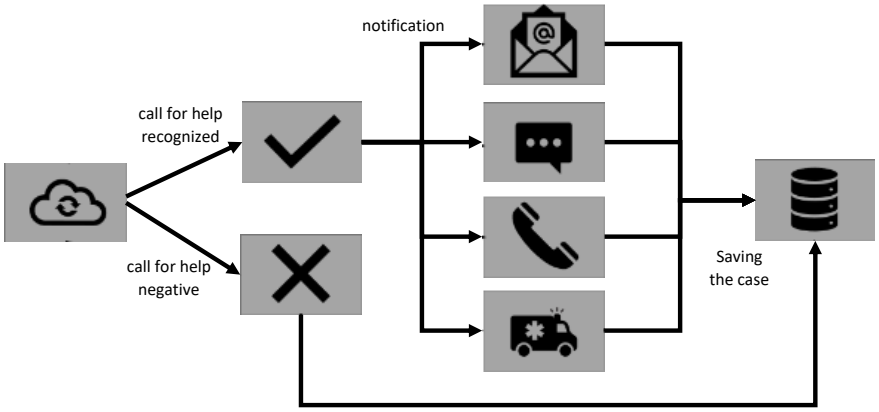

Fig. 4. Notification and saving the information on the recognized intention.

4 Experimental Results

Appropriate reaction to the current situation requires correct recognition of the intention that is hidden in the voice command. The quality of intention recognition depends on the effectiveness of both steps that are performed during voice analysis. Here, we compare various Cloud-based services for speech recognition and natural language understanding in terms of their capabilities to serve the purpose of the designed system.

4.1 Quality Assessment for Speech to Text Transcription

Relying on integrated platforms with well-tested services is one of the most common ways developers build their solutions that cover many loosely coupled components (including IoT devices for personal assistance and their software). When designing our Cloud-based systems to call for help and monitor older adults, we also followed this pattern. Therefore, when testing the speech-to-text transcription services, we mainly looked at large Cloud platforms that provide the possibility to connect many personal IoT devices and integrate them within a coherent system. For the speech to text transcription, we tested the following services:

– Google Cloud Speech-to-Text,

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77970-2_36

– IBM Watson Speech to Text,
– Microsoft Azure Speech to Text,
– Amazon Transcribe.

The research was carried out on a VoxForge data set[4] consisting of 551 files in WAV format with different sampling rates (16 kHz or 44kHz to 48kHz). This collection was selected from speech recordings in the English language. Speech recordings are of various dialects: British English, European and American. The length of the recordings was usually from 4 to 10 seconds, which was adapted to the length of the danger notification commands. The test data set consisted of recordings (WAV, waveform audio format) and text files containing the correct transcripts corresponding to the given WAV file. After obtaining a result in the form of a text sentence for a given cloud service, it was compared with the pattern (correct transcript, a ground truth).

One of the most popular ways to measure the quality of speech recognition is to evaluate effectiveness using the *Word error rate* (WER) [15]. WER is a simple index, which is the quotient of the sum of substituted S, omitted D and inserted I words by their number in a given sentence N. One of the variants of this metric is to use a weight of 0.5 for the removed and inserted words:

$$WER = \frac{S + D + I}{N}. \tag{1}$$

The second metric used to examine the entire group of results of a given service is *Sentence Error Rate* (SER) – determining the sentence recognition error rate. It is the number of incorrectly recognized sentences F in a given research group N:

$$SER = \frac{F}{N}. \tag{2}$$

Values of the *Word error rate* (WER) achieved in our experiments for various speech-to-text cloud services are presented in Fig. 5. Values of the WER for all tested services are low, indicating the possible usefulness of each of the services for the task being performed. The lowest value of $WER = 8.40$ was achieved with the Amazon Transcribe service. The largest value of $WER = 13.80$ was achieved for the IBM Watson Speech to Text.

Values of the *Sentence Error Rate* (SER) achieved in our experiments for various speech-to-text cloud services are presented in Fig. 6. Values of the SER for all tested services are also low, which confirms the possible usefulness of each of the services for the task being performed. The lowest value of $SER = 4.05$ was achieved with the Microsoft Azure Speech to Text service.

[4] VoxForge open speech dataset with transcribed speech: `http://www.voxforge.org/home/downloads/speech/english`

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77970-2_36

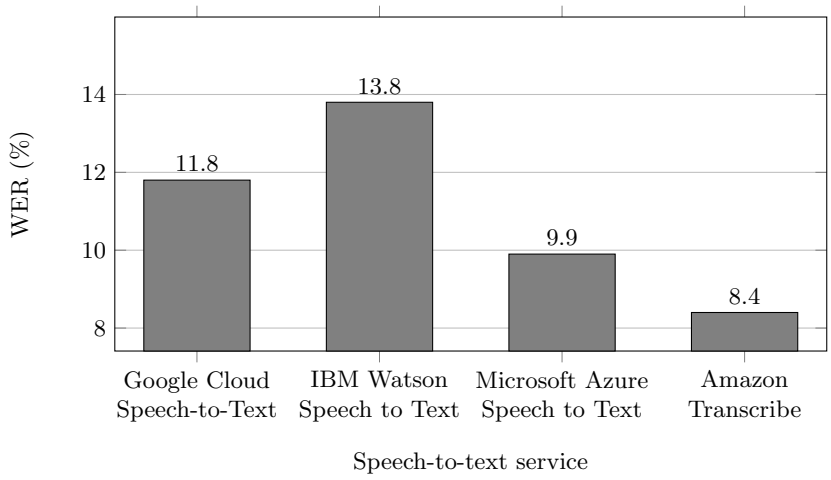

Fig. 5. Values of the *Word error rate* (WER) achieved for various speech-to-text cloud services.

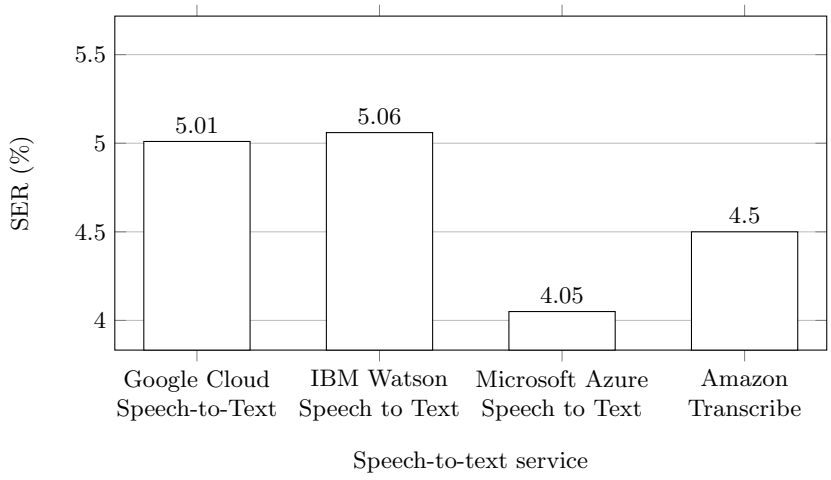

Fig. 6. Values of the *Sentence Error Rate* (SER) achieved for various speech-to-text cloud services.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77970-2_36

4.2 Experimental Evaluation of Natural Language Understanding Services

In the second series of experiments, we conducted tests of the capabilities of Cloud services to understand natural language. We investigated Google DialogFlow, IBM Watson Assistant, Amazon Lex (the same deep learning technologies that power Amazon Alexa), Microsoft LUIS, and the free Rasa NLU library for this purpose. Within these experiments, we wanted to:

- verify the effectiveness of speech understanding depending on the service used,
- study the impact of the size of the training set on the effectiveness of the speech understanding services in use.

As in the case of speech recognition, we could not find any test data related to the Emergency Command issues. Alternatively, for our studies, we used the SNIPS dataset [7], consisting of 7 intentions:

- GetWeather – related to weather conditions,
- BookRestaurant – related to booking a meal in a given restaurant,
- PlayMusic – related to playing an artist, album or music track,
- AddToPlaylist – related to adding a music track to a playlist,
- RateBook - related to book reviews,
- SearchScreeningEvent – related to searching for film events,
- SearchCreativeWork – related to searching for creative activities.

The full data set used in our experiments contained 13,784 sentences (approximately 2,000 sentences for each intention), and besides, each file with intentions contained between 3,419 and 6,418 entities. We used two data sets for the training process: (1) a limited data set that contained 300 samples per intention, and (2) full data set with approximately 2,000 sentences per intention. For the limited training set, the number of entities ranged from 533 to 1,146. The test set consisted of 700 sentences (100 for each intention). Importantly, the test set for each intention was already distinguished. Therefore, no cross-validation or other modifications of the training set were required. Table 1 contains results of effectiveness evaluation for various NLU services trained with up to 300 samples per intention. Table 2 contains results of effectiveness evaluation for various NLU services trained with around 2000 samples per intention.

The largest number of properly-identified intentions was achieved with Microsoft LUIS service - 99.14% and 98.57% - and Rasa NLU - 98.57% for both data sets. For most of the services, we noticed an increase in effectiveness measures for the larger data set used for the training phase. This was expected. However, it is interesting that despite the increase in the number of training data, Microsoft LUIS was the only one that experienced a decrease in effectiveness. In the prediction with models trained with the limited data set, Microsoft LUIS made only two mistakes in the intention recognition for the testing sentences (Table 3). For the prediction based on training with the full data set, Microsoft LUIS made 10 wrong predictions (Table 4).

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77970-2_36

Table 1. Effectiveness of NLU for various services trained with the limited data set of 300 samples per intention.

NLU Service	Recognized intentions (%)	F1	Precision	Recall	Recognized entities (%)	Mean time (ms)
Amazon Lex	93.71	0.5582	0.5876	0.5741	50.01	277.88
Google DialogFlow	96.71	0.6022	0.7522	0.5266	55.17	328.95
Microsoft LUIS	99.14	0.4409	0.4626	0.4739	56.45	202.92
Rasa NLU	98.57	0.6526	0.6705	0.6446	80.64	9.50
IBM Watson NLU	98.14	0.4987	0.5130	0.5369	56.71	289.70

Table 2. Effectiveness of NLU for various services trained with the full data set of 2000 samples per intention.

NLU Service	Recognized intentions (%)	F1	Precision	Recall	Recognized entities (%)	Mean time (ms)
Amazon Lex	94.14	0.6089	0.6207	0.6347	56.83	286.87
Google DialogFlow	98.14	0.6888	0.7797	0.6393	67.94	337.76
Microsoft LUIS	98.57	0.5594	0.5434	0.6135	77.57	113.89
Rasa NLU	98.57	0.6794	0.6871	0.6766	86.42	9.56
IBM Watson NLU	98.43	0.5152	0.4791	0.6225	69.32	280.34

Table 3. Sentences with wrongly assigned intentions by Microsoft LUIS trained with the limited data set.

Sentence	Real intention	Predicted intention
When is sunrise for AR	GetWeather	SearchScreeningEvent
I want to eat in Ramona	BookRestaurant	GetWeather

Table 4. Sentences with wrongly assigned intentions by Microsoft LUIS trained with the full data set.

Sentence	Real intention	Predicted intention
When is sunrise for AR	GetWeather	SearchScreeningEvent
I want to eat in Ramona	BookRestaurant	GetWeather
Live In L.a Joseph Meyer please	PlayMusic	BookRestaurant
Where is Belgium located	GetWeather	SearchCreativeWork
Please tune into Chieko Ochi 's good music	PlayMusic	AddToPlaylist
I want to see JLA Adventures: Trapped In Time	SearchScreeningEvent	SearchCreativeWork
Where can I see The Prime Ministers: The Pioneers	SearchScreeningEvent	SearchCreativeWork
I want to see Medal for the General	SearchScreeningEvent	SearchCreativeWork
I want to see Fear Chamber	SearchScreeningEvent	SearchCreativeWork
I want to see Outcast	SearchScreeningEvent	SearchCreativeWork

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77970-2_36

4.3 Influence of the manner of speaking and noise on the quality of recognition of emergency calls

The third part of the experiments covered investigations of the influence of the manner of speaking and the impact of noise on the quality of recognition of emergency calls. For this purpose, we recorded 15 sentences (including 13 calls for help and two random commands) spoken by each person. The study aimed to check how the emergency commands detection system reacts in case of danger and how susceptible it is to the change in the way of speaking and the surrounding noise. Tests covered four manners of pronouncing calls for help: normal speech, whispering, slurred speech, and calling for help with background noise. We performed the tests with the Microsoft LUIS service, which achieved the best results in the previous series of intention recognition tests. Results are presented in Fig. 7.

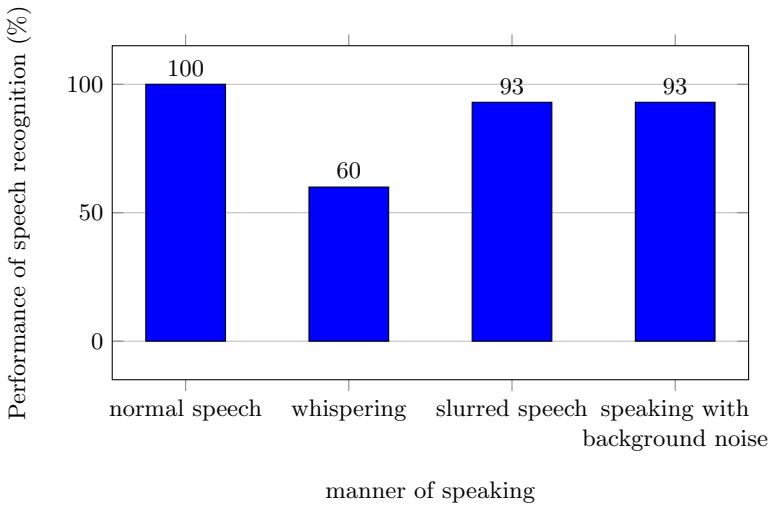

Fig. 7. Effectiveness (%) in recognizing intentions in uttered calls for help depending on the manner of speaking. Obtained for Microsoft LUIS service.

As can be observed from these experiments, the best results were obtained for normal speaking – the effectiveness reached 100%. Very good results were also obtained for slurred speech and speaking with a background noise – both 93%. The worst results were obtained for whispering, for which we achieved only 60% of properly-recognized intentions. This is an important observation, since after an accident older people may not be able to speak normally. This leaves room for future research on the development of algorithms for NLU for these types of situations.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77970-2_36

5 Discussion and Conclusions

In the conducted study, we compared the capabilities of speech recognition and natural language understanding services. In the evaluation of speech recognition services, two services turned out to be the best – Amazon Transcribe with $WER = 8.4\%$ and $SER = 4.50\%$ and Microsoft Azure Speech to Text with $WER = 9.90\%$ and $SER = 4.05\%$. This shows that both cloud services are among the leaders in this field. The obtained values of the WER error are several times lower than in related works [28], which may indicate that we used a simpler test set in our experiments. The speech recognition execution time is at the level of a few seconds, which allows using these methods (that require Cloud communication) in emergency call applications.

In the second part of the experiments, we compared the results of speech understanding for Microsoft LUIS, Google DialogFlow, Amazon Lex, IBM Watson Assistant, and Rasa NLU services to recognize the intention of the speech. When trained with the full data set (about 2,000 sentences per intention), Microsoft LUIS and Rasa NLU turned out to return the best quality results. For services trained with the limited data set (300 sentences per intention), Microsoft LUIS turned out to be the most effective. These results and the order of services in terms of top-performing ones correspond well with related studies [6].

Comparing the effectiveness of entity recognition (intention discovery) by using the F1 indicator, we can conclude that among the tested services, the most efficient was Rasa NLU for the model trained with the limited data set – $F1 = 0.6526$ and Google DialogFlow for the full data set - $F1 = 0.6888$. Considering the percentage of the recognized intentions, Microsoft LUIS proved to be the most effective, with 99.14% for the model trained with the limited data set and 98.57 for the model trained with the full data set. Anyway, all tested services achieved a recognition rate above 93%, which is a good result.

In summary, it is worth mentioning that Rasa NLU achieved excellent results compared to paid Cloud-based services, being in the lead when recognizing speech intention in each of the categories. Among Cloud-based services, excellent results were achieved by services offered by Microsoft, which allow for the inclusion of these services in the construction of emergency call applications. In the case of speech understanding, comparing our results to related works, we can conclude that the key is to properly adjust the training data. For each intention, statements should be thoughtful and as unique as possible compared to other intentions. For example, in detecting dangers in the health state of older people, it would be reasonable to train the people to start the emergency call with some predefined expression, like 'Emergency, I need help.' It would also protect against an unintentional analysis of a speech related to everyday life and human conversations (i.e., reduce the amount of data that undergoes the analysis). This is also important from the viewpoint of older adults' privacy and information security since older people want to be sure that they are not overheard by IoT devices, other people, and unauthorized third parties. Therefore, data security and peoples' privacy are important issues when using IoT solutions that serve people.

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77970-2_36

References

1. World Health Organization: Global health and aging. Tech. Rep. 11-7737, NIH Publication (2011)
2. Ammari, T., Kaye, J., Tsai, J.Y., Bentley, F.: Music, search, and IoT: How people (really) use voice assistants. ACM Trans. Comput.-Hum. Interact. **26**(3) (2019)
3. Austerjost, J., Porr, M., Riedel, N., Geier, D., Becker, T., Scheper, T., Marquard, D., Lindner, P., Beutel, S.: Introducing a virtual assistant to the lab: A voice user interface for the intuitive control of laboratory instruments. SLAS Technology: Translating Life Sciences Innovation **23**(5), 476–482 (2018)
4. Bhosale, S., Sheikh, I., Dumpala, S.H., Kopparapu, S.K.: Transfer learning for low resource spoken language understanding without speech-to-text. In: 2019 IEEE Bombay Section Signature Conference (IBSSC). pp. 1–5 (2019)
5. Braines, D., O'Leary, N., Thomas, A., Harborne, D., Preece, A.D., Webberley, W.M.: Conversational homes: a uniform natural language approach for collaboration among humans and devices. Int. J. Intell. Syst. **10**(3), 223 – 237 (2017)
6. Braun, D., Hernandez Mendez, A., Matthes, F., Langen, M.: Evaluating natural language understanding services for conversational question answering systems. In: Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue. pp. 174–185. Association for Computational Linguistics, Saarbrücken, Germany (2017)
7. Coucke, A., Saade, A., Ball, A., Bluche, T., Caulier, A., Leroy, D., Doumouro, C., Gisselbrecht, T., Caltagirone, F., Lavril, T., Primet, M., Dureau, J.: Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. ArXiv **abs/1805.10190** (2018)
8. Cupek, R., Drewniak, M., Fojcik, M., Kyrkjebø, E., Lin, J.C.W., Mrozek, D., Øvsthus, K., Ziebinski, A.: Autonomous guided vehicles for smart industries – the state-of-the-art and research challenges. In: Krzhizhanovskaya, V.V., Závodszky, G., Lees, M.H., Dongarra, J.J., Sloot, P.M.A., Brissos, S., Teixeira, J. (eds.) Computational Science – ICCS 2020. pp. 330–343. Springer International Publishing, Cham (2020)
9. de Velasco, M., Justo, R., Antón, J., Carrilero, M., Torres, M.I.: Emotion Detection from Speech and Text. In: Proc. IberSPEECH 2018. pp. 68–71 (2018)
10. Deng, L., Li, J., Huang, J., Yao, K., Yu, D., Seide, F., Seltzer, M., Zweig, G., He, X., Williams, J., Gong, Y., Acero, A.: Recent advances in deep learning for speech research at microsoft. In: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. pp. 8604–8608 (2013)
11. Dimauro, G., Di Nicola, V., Bevilacqua, V., Caivano, D., Girardi, F.: Assessment of speech intelligibility in parkinson's disease using a speech-to-text system. IEEE Access **5**, 22199–22208 (2017). https://doi.org/10.1109/ACCESS.2017.2762475
12. Graves, A., Fernández, S., Gomez, F., Schmidhuber, J.: Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In: Proceedings of the 23rd International Conference on Machine Learning. p. 369–376. ICML '06, Association for Computing Machinery, New York, NY, USA (2006). https://doi.org/10.1145/1143844.1143891
13. Grzechca, D., Ziebinski, A., Rybka, P.: Enhanced reliability of ADAS sensors based on the observation of the power supply current and neural network application. In: Nguyen, N.T., Papadopoulos, G.A., Jedrzejowicz, P., Trawiński, B., Vossen, G. (eds.) Computational Collective Intelligence. pp. 215–226. Springer International Publishing, Cham (2017)

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77970-2_36

14. Kishore Kodali, R., Rajanarayanan, S.C., Boppana, L., Sharma, S., Kumar, A.: Low cost smart home automation system using smart phone. In: 2019 IEEE R10 Humanitarian Technology Conference (R10-HTC)(47129). pp. 120–125 (2019)
15. Klakow, D., Peters, J.: Testing the correlation of word error rate and perplexity. Speech Communication **38**(1), 19 – 28 (2002)
16. Lago, A.S., Dias, J.P., Ferreira, H.S.: Conversational interface for managing non-trivial internet-of-things systems. In: Krzhizhanovskaya, V.V., Závodszky, G., Lees, M.H., Dongarra, J.J., Sloot, P.M.A., Brissos, S., Teixeira, J. (eds.) Computational Science – ICCS 2020. pp. 384–397. Springer International Publishing, Cham (2020)
17. Laksono, T.P., Hidayatullah, A.F., Ratnasari, C.I.: Speech to text of patient complaints for bahasa indonesia. In: 2018 International Conference on Asian Language Processing (IALP). pp. 79–84 (2018). https://doi.org/10.1109/IALP.2018.8629161
18. Lero, R.D., Exton, C., Le Gear, A.: Communications using a speech-to-text-to-speech pipeline. In: 2019 International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob). pp. 1–6 (2019)
19. López, G., Quesada, L., Guerrero, L.A.: Alexa vs. siri vs. cortana vs. google assistant: A comparison of speech-based natural user interfaces. In: Nunes, I.L. (ed.) Advances in Human Factors and Systems Interaction. pp. 241–250. Springer International Publishing, Cham (2018)
20. Mehrabani, M., Bangalore, S., Stern, B.: Personalized speech recognition for internet of things. In: 2015 IEEE 2nd World Forum on Internet of Things (WF-IoT). pp. 369–374 (2015). https://doi.org/10.1109/WF-IoT.2015.7389082
21. Mishakova, A., Portet, F., Desot, T., Vacher, M.: Learning natural language understanding systems from unaligned labels for voice command in smart homes. In: 2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops). pp. 832–837 (2019)
22. Mishakova, A., Portet, F., Desot, T., Vacher, M.: Learning natural language understanding systems from unaligned labels for voice command in smart homes. In: 2019 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops). pp. 832–837 (2019)
23. Mitrevski, M.: Conversational Interface Challenges, pp. 217–228. Apress, Berkeley, CA (2018)
24. Mrozek, D., Koczur, A., Małysiak-Mrozek, B.: Fall detection in older adults with mobile IoT devices and machine learning in the cloud and on the edge. Information Sciences **537**, 132 – 147 (2020)
25. Mrozek, D., Milik, M., Małysiak-Mrozek, B., Tokarz, K., Duszenko, A., Kozielski, S.: Fuzzy intelligence in monitoring older adults with wearables. In: Krzhizhanovskaya, V.V., Závodszky, G., Lees, M.H., Dongarra, J.J., Sloot, P.M.A., Brissos, S., Teixeira, J. (eds.) Computational Science – ICCS 2020. pp. 288–301. Springer International Publishing, Cham (2020)
26. Schwitter, R.: Controlled natural languages for knowledge representation. vol. 2, pp. 1113–1121 (01 2010)
27. Sovariova Soosova, M.: Determinants of quality of life in the elderly. Central European Journal of Nursing and Midwifery **7**(3), 484–493 (2016)
28. Vyas, M.: A Gaussian mixture model based speech recognition system using Matlab. Signal & Image Processing **4**(4), 109 – 118 (2013)
29. Wan, J., A. A. H. Al-awlaqi, M., Li, M., O'Grady, M., Gu, X., Wang, J., Cao, N.: Wearable iot enabled real-time health monitoring system. EURASIP Journal on Wireless Communications and Networking **2018**(1), 298 (Dec 2018)

ICCS Camera Ready Version 2021
To cite this paper please use the final published version:
DOI: 10.1007/978-3-030-77970-2_36