Automatic Help Summoning through Speech Analysis on Mobile Devices

Bożena Małysiak-Mrozek[1], Paweł Wojaczek[2], Krzysztof Tokarz[3], Vaidy Sunderam[4], Dariusz Mrozek[2], and Jean-Charles Lamirel[5]

[1] Department of Distributed Systems and Informatic Devices, Silesian University of Technology, Akademicka 16, 44-100 Gliwice, Poland
[2] Department of Applied Informatics, Silesian University of Technology Akademicka 16, 44-100 Gliwice, Poland
[3] Department of Graphics, Computer Vision and Digital Systems, Silesian University of Technology, Akademicka 16, 44-100 Gliwice, Poland
[4] Department of Computer Science, Emory University, Atlanta, GA 30322, USA
[5] Synalp team, LORIA, Nancy - University of Strasbourg, Strasbourg, Grand Est, France
dariusz.mrozek@polsl.pl

Abstract. The aging population has led to an increasing number of elderly individuals living alone, making it crucial to address their need for prompt and effective emergency assistance. Older adults, often facing physical limitations or illnesses, require reliable systems for immediate help during life-threatening situations. To meet this need, smart devices like emergency call systems are being developed, enhancing seniors' safety and improving health and social care responses. Our research explores how passive and active speech analysis on mobile devices can support automatic emergency assistance. We show that this can be achieved on Edge devices using tiny machine learning (ML) models for wake-word detection, speech-to-text conversion, and intention recognition, paving the way for safer, smarter living environments for seniors.

Keywords: Internet of Things, older adults, Natural Language Processing, intention recognition, speech analysis

1 Introduction

The aging society, characterized by a growing proportion of older individuals in the population, is a global phenomenon [18]. Advances in healthcare have improved living conditions, enabling more people to reach old age [1]. For the first time, projections suggest that the global population over 65 could soon exceed the number of children under 5 [26]. This trend spans all regions worldwide [18].

In 2021, 761 million people worldwide were over 65, including 155 million aged 80 or older (Table 1). By 2050, these numbers are expected to rise to 1.6 billion and 459 million, respectively. Older individuals constituted 9% of the global population in 2021, a figure projected to reach 16% by 2050—meaning

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97573-8_2

Table 1: The number of older people in the global society in 2021 and 2050 [28].

Year	2021	2050
The number of people over 65 years old	761 million	1.6 billion
The number of people over 80 years old	155 million	459 million
Percentage of people over 65 years old in total population	9%	16%

one in six people will be elderly. Additionally, life expectancy for those over 65 is expected to increase from 82 years in 2021 to 84 years by 2050 in more developed regions [28]. Older people who live alone form a social group that requires special attention, support, and care from others due to their health and physical conditions. It may seem natural that this care can be provided by their relatives, such as children or grandchildren, or if they do not have any, siblings, or people from extended family. Statistical analyses have shown that, for the most part, even when adult children no longer live with their parents, they still live close to their place of residence, which certainly makes it easier for them to provide help to their parents [27]. Elderly people can also benefit from the help of employed home caregivers. Various types of social services, helping the elderly, charities, and self-help groups, are also being established.

Nevertheless, the above-mentioned forms of assistance are often not practiced continuously at any time of the day or night. Therefore, devices that enable reliably calling for help at any time are needed, especially in situations of immediate life-threatening risk caused by illness or accident. These may be electronic devices, edge Internet of Things (IoT) solutions, or smartphones that enable remote notification of appropriate people or services about the need to provide help. The introduction of these solutions becomes a key element in caring for the health of seniors, providing support for their families and caregivers, and contributing to an improvement in the quality of life within an aging society.

This paper shows how the automatic call for help (ACH) can be initialized through passive and active voice analysis on a mobile device. Section 3 discloses the general idea and architecture of ACH with voice analysis models. Section 4 describes the algorithm designed to perform this process, and Section 5 explains the created experimental environment with the ML models used for particular phases of ACH. In Section 6, we experimentally test wake-up keyword detection and voice-to-text transcription for various noise conditions. We also verify the effectiveness of different ML-based intention recognition models operating at edge devices and those provided as web services. Our experiments confirm that ACH can be efficiently performed with tiny ML models designed for smartphones.

2 Related works

In today's society, where caring for the elderly and disabled plays a key role, the use of smart technologies to create devices that enable calling for help has become indispensable. These devices are intended to ensure a quick response in emergency situations and provide a sense of security for both people in need

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97573-8_2

of support and their caregivers [3]. The number of devices was developed and described in the literature. They are usually equipped with a wearable IoT device with a set of sensors and a stationary receiver with internet connectivity. The stationary device can be a mobile phone, like in PhystioDroid created by Banos et al. [7], or a separate device, like in the system developed by Lersilp et al. [17]. The carried device can have a form of the wristband [17], the belt mounted at the waist level [30], or chest level [7]. Sensors can be integrated with the wheelchair [14] or with the user's clothes [13]. Wearable devices contain mainly sensors for body temperature, movements, heart rate, blood pressure and haemoglobin oxygen saturation SpO2. Critical features are the ability of the fall detection [23], [16] and the alarm trigger, which can have a form of a large red SOS button [17], [2]. In a more advanced system, the emergency situation can be cancelled by a voice command or confirmed by the command or lack of it, causing the notification of relatives or medical services.

There is a group of systems relying on speech recognition technology, which makes it easier to operate with, especially for people with lower digital technology skills, since speech is the most natural way of communicating between people. Such systems usually consist of two main modules - Automatic Speech Recognition (ASR) and then Natural Language Understanding (NLU) [8]. ASR systems can recognise separate words, while more advanced NLU systems try to understand their meaning when spoken in conjunction with other words. NLU is still a complex task for computers [9], especially in an emergency situation, when patients may not be able to speak clearly. Additional information can be obtained with emotion detection [11]. The availability of high computing power and artificial intelligence (AI) services in computing clouds has enabled the creation of voice recognition systems in the form of voice assistants. These are currently Apple's Siri [6], Microsoft Cortana (2014), Amazon Alexa (2014), and Google Assistant (2016). One of the newest achievements in this area is the OpenAI Whisper [24]. Research shows that it can even successfully recognise the speech for dysarthric patients [29]. Currently, the NLP in the cloud is more effective than voice recognition technology built into end devices [6], but constant development and miniaturisation of microcontrollers enable progress in voice recognition on the Edge. In 2021, Mrozek et al. [20] also investigated ASR approaches available in the cloud. However, their usage was highly dependent on the availability of cloud services and stable Internet access. The current work describes the system that moves this process to the Edge, which is essential in situations of lack or poor Internet connection. However, we also compare the Edge-based models with the external, cloud-based ones.

3 General Idea of Calling for Help

Our idea for calling for help in dangerous situations that may occur in seniors' lives assumes the use of a smartphone to monitor the surrounding speech in standby mode and wake up and react only when a monitored person (senior) utters a call for help. Although not every senior may have one, smartphones are

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97573-8_2

increasingly used by this group of people. As electronic devices, today's smartphones not only provide connectivity in terms of voice and data transmission but are also powerful enough to perform sophisticated computational operations, such as data analysis involving ML-based inference. This analysis may cover ASR and NLU, as shown in Fig. 1. Some of these processes can also be optionally implemented by invoking external services. However, such a flow requires a stable internet connection, additional GSM data transmission from a mobile device (such as a smartphone) to these services, and the constant availability of external services. Therefore, local, edge-based help summoning should work more efficiently and support real-time reactions in the event of danger. This will be verified experimentally (Section 6). In an emergency, the smartphone calls the caregiver or informs him of what happened. Another argument for using smartphones is that they can detect dangerous situations (e.g., falls) automatically based on built-in IMU sensors. This enables the integration of various monitoring approaches into a single mobile device, suitable for use in both smart home environments and outdoors in smarter cities.

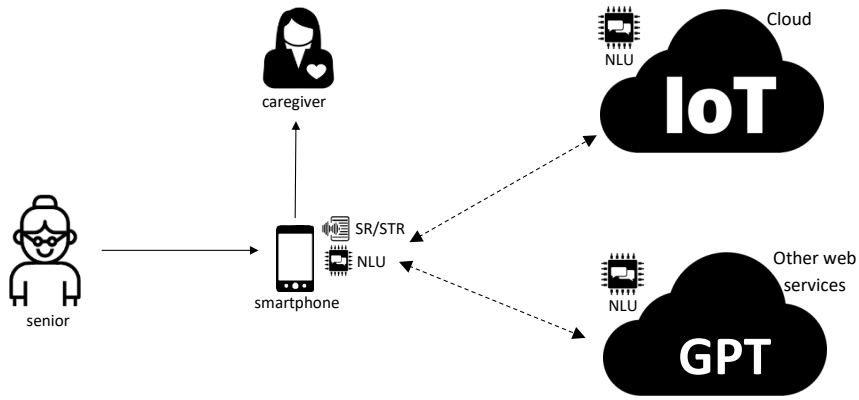

Fig. 1: General architecture of the senior monitoring environment with automatic help summoning through speech analysis on a mobile device.

4 Speech Stream Processing for Help Summoning

Every day, people speak on many topics, formulating sentences consisting of words that create a logical sequence of statements. Each statement can, therefore, be treated as a stream of sentences and the words that compose them. Let S be a statement stream with n sentences s spoken by a person:

$$S = \{s_i | i = 1 \dots n, n \in \mathbb{N}^+\}. \tag{1}$$

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97573-8_2

Each sentence $s_i \in S$ consists of a different number of words w:

$$s_i =< w_1, w_2, \ldots, w_m >, \tag{2}$$

where $w_1, w_2, \ldots, w_m$ are particular words and m is a (varying) number of words in a sentence.

Calling for help using smart devices that analyze speech consists of several stages. The first one assumes that the intention to call for help appears after saying a special wake-up keyword w_k.

$$s_i =< w_1, w_2, \ldots, [w_k], \ldots, w_m >, \tag{3}$$

where $[w_k]$ denotes that the keyword may appear optionally in a sentence. However, once it appears as a spoken word, it starts the second and the third stages of the analysis for the rest of the statement stream:

$$S' = \{< w_k, w_{k+1} \ldots, w_m >, s_{i+1}, \ldots s_n\}. \tag{4}$$

The process is, therefore, carried out in three stages, as shown in Fig. 2. In the idle state, the device listens for the user to say a defined keyword, monitoring the signal from the device's built-in microphone. For the purposes of this work, we chose the word *help*. It is easily recognized, short, characterized by clear sound, and commonly used in situations requiring urgent intervention or support. After detecting the utterance of a keyword, the application enters the active state in which it transcribes the words spoken by the user, transforming speech into text.

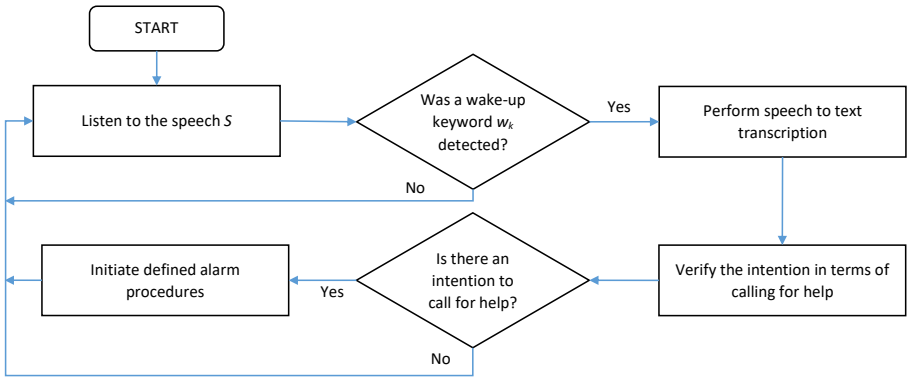

Fig. 2: General algorithm of automatic help summoning through speech analysis on a mobile device.

In such a form, the transcribed speech can be classified more effectively using computer technology. Transcription continues until the user stops speaking any more words for a defined period. The final stage covers the classification of the user's statements to recognize the intention to call for help. The application

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97573-8_2

determines whether the user had actual intention to call for help or whether his statement was accidental or unrelated to an emergency situation. The classification process relies on analyzing the content of the transcript and defining the probability of an alarm situation occurring. If a true intention to call for help is detected, the device is ready to initiate defined alarm procedures, e.g., to inform the appropriate emergency services about the location and nature of the reported event or to record this fact in a database available to the user's caregivers. Otherwise, the device remains idle, ready for any further sound signals.

5 Experimental Environment and Methods

To investigate the automatic call for help, we developed a mobile application for a smartphone (Dart programming language and Flutter framework) with a backend layer implementing the speech analysis algorithm presented in Fig. 2 and a frontend to visually observe the results of the analysis.

For keyword detection, we considered several solutions, including Pocketsphinx [15], Mycroft Precise, Snowboy, and Porcupine Wake Word by Picovoice, and finally selected and tested the last one as it fitted our requirements. The capability of detecting the wake word to stitch into active mode is particularly valuable in voice-controlled applications and devices, offering a seamless and efficient means of interaction [21]. The Porcupine supports multiple languages, many target platforms, operating systems, and programming languages, which allows building models optimized for a given device. In our case, we created an Android model, and we trained it to detect the wake-up keyword *help*. Then, we could process voice data fully locally on the device without the use of external servers or services. This locality eliminated the impact of network delays, access interruptions, or bandwidth limitations on the quality of voice analysis.

For the second stage of the speech analysis, i.e., speech-to-text transcription, we considered several, mainly open-source toolkits and libraries, including Kaldi [22], DeepSpeech engine by Mozilla [12], Pocketsphinx [15], and Vosk [4]. Based on the comparisons of the quality and performance reported in [25], we focused on Kaldi and Vosk and finally chose Vosk, which supports over 20 languages and allows building small models (approx. 50MB) intended for use on smartphones and single-board computers such as Raspberry Pi. They require relatively little computing power and memory to operate [5].

For the fundamental stage of detecting the intention to call for help in the user's statements, we tested six analytical solutions. For local intent detection, we created three different models using the TensorFlow Lite (TFLite) library. Additionally, we created a fourth model using a light embedding approach. These models have been optimized to operate on devices with limited resources and low energy consumption, such as portable devices, smartphones, microcontrollers, and embedded systems.

The first model (*TF simple*) is a simple sequential model acting as a classifier composed of several layers (Fig. 3a). The initial layer is the *TextVectorization*, tasked with transforming the text into a sequence of token indices. The next

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97573-8_2

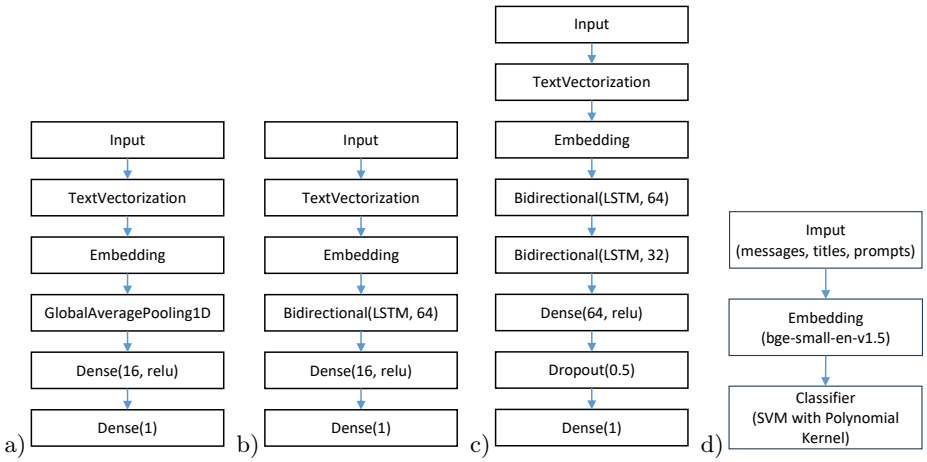

Fig. 3: Local inference models for intent detection: (a) simple, (b) 1-layer Bidirectional LSTM, (c) 2-layer Bidirectional LSTM, and (d) Embed&Class.

one, the *Embedding* layer, holds a vector for each word. Upon invocation, it transforms sequences of word indices into sequences of vectors. The *GlobalAveragePooling1D* layer produces a constant-length output vector for each example by computing the average across the sequence dimension. This approach allows the model to manage variable-length input data in a straightforward manner, which is crucial given the varying number of words in input statements. The constant-length output vector is passed through a fully connected *Dense* neural network layer with 16 hidden units. The final *Dense* layer is fully connected with a single output node. The second model (*TF 1-layer Bi-LSTM*, Fig. 3b) is slightly more complicated and uses the bidirectional LSTM recurrent neural network and fixed-length representation of text in the *Embedding*. The third model (*TF 2-layer Bi-LSTM*, Fig. 3c) differs from the second one by adding one more recurrent bidirectional LSTM layer directly after the first existing bidirectional LSTM layer and the *Dropout* layer, which helps prevent overfitting. The fourth model, we named Embed&Class, operates in a different way. In the first, embeddings are produced by the use of the Qdrant FastEmbed library [6], a library designed for fast embedding and which can rely on lightweight quantized language models based on Transformers, like bge-small-en-v1.5, a small-scale English text embedding model developed by BAAI (Beijing Academy of Artificial Intelligence). This approach is then combined with a standard SVM classifier that operates directly on the generated embedding vectors to detect intentions. The SVM model is trained using a 10-fold cross-validation process. Input embedding vectors used for learning are normalized. As parameters, the SVM uses a ridge logistic regressor for data calibration, a polynomial kernel, and a C value of 1.

[6] https://github.com/qdrant/fastembed

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97573-8_2

As an external tool for intent recognition and the fifth analytical solution, we tested Azure Cognitive Services. It is a set of cloud services available in Microsoft Azure that enable the integration of artificial intelligence and machine learning across applications. These services offer pre-trained models for processing text, video, and speech data, as well as services for analyzing geospatial data. The default, pre-trained model has been tuned by us to the problem of calling for help. Specifically, the model was tuned to assign the sentence to one of two classes - the class of sentences containing or not the intention of calling for help.

As the last analytical approach, we also tested the external GPT-3.5 (Generative Pre-trained Transformer 3.5) service by OpenAI. It provided a stable (at the time of system implementation) and advanced language model based on the Transformer architecture, a neural network specifically designed to process sequences of data, such as text.

6 Experiments

All three stages of the speech analysis were experimentally tested on Samsung S20 with Android 13 using the implemented mobile application to assess the feasibility of performing automated calling for help at the Edge.

6.1 Wake-up Keyword Detection

To assess the quality of wake-up word detection, we utilized a dataset comprising 300 voice recordings in WAV format, with a sampling frequency of 16 kHz. The dataset was balanced, comprising 150 speech recordings with the *help* wake-up word among other sentences spoken, and 150 recordings without the wake-up word. The dataset was created based on sentences from the Kaggle *Medical Speech, Transcription, and Intent* dataset [19] with audio statements related to typical medical symptoms along with their transcriptions. Since this dataset does not contain any sentences with the wake-up word, based on its content, we recorded 300 statements with ten 60+ volunteers speaking 30 selected sentences that contained or did not contain the wake-up word from a distance of 40 cm.

The detection process was treated as a two-class classification problem, where the output of the detector was a logical value indicating whether it recognized the keyword in the recording or not. Based on the obtained confusion matrix, where a *True Positive* was a correctly detected recording with the wake-up word, while the *True Negative* was a correctly detected recording without a wake-up word, we calculated the values of performance metrics, including accuracy, precision, recall, and F1-score (Table 2). The research was carried out on original recordings and the recordings with white noise added. The maximum amplitude of the recording was assumed as the default value for the maximum noise amplitude. We also tested the recordings with added noise of the amplitudes reduced by 20, 40, and 60 dB compared to the default value.

As can be observed, the accuracy of detecting the wake-up word was the highest for the original sound and is as high as 98%. As white noise is added,

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97573-8_2

Table 2: Performance of the wake-up word detection

	Accuracy	Precision	Recall	F1-score
Original sound	0.980	0.993	0.967	0.980
With white noise (-60 dB)	0.963	1.000	0.927	0.962
With white noise (-40 dB)	0.943	1.000	0.887	0.940
With white noise (-20 dB)	0.827	1.000	0.653	0.790
With white noise (-0 dB)	0.500	N/A	0.000	N/A

the accuracy decreases, but for noise with an amplitude reduced by 60 and 40 dB, the detector still achieves satisfactory results above 94%. Adding noise with an amplitude reduced by 20 dB causes a significant deterioration of the model's accuracy, which drops to 82.7%. The effect of adding noise with the default amplitude is the complete cessation of detection of the wake-up word, as evidenced by an accuracy value of 50% combined with a sensitivity (recall) of zero. Precision values of 1 in most analyzed cases suggest that the model is never triggered by phrases not containing the *help* word. The exception is only one case for the original sound, which contributed to the precision value in this case being 0.993. However, this is still a very high value. The key metric in the context of the considered application of the detector is sensitivity (recall). It tells us what proportion of real distress calls were correctly detected. For the original sound, the value of this metric reached 0.967. This is a satisfactory result. The sensitivity for recordings with added noise is slightly worse. However, for a noise amplitude reduced by 60 dB, the model still achieves a value exceeding 90%. As the amplitude of the added noise increases, the sensitivity value decreases, and for noise with an amplitude reduced by 20 dB, the software detects only about two-thirds of the distress calls. Adding noise at the default high amplitude means that the wake-up word is not detected in any of the test recordings, as evidenced by a sensitivity value of zero. The values of the F1-score, which is the harmonic mean of precision and sensitivity, also show similar behavior and decrease as the amplitude of the added noise increases. However, it is worth noting that the noise values corresponding to the two highest considered amplitudes are rarely encountered in real-life conditions.

6.2 Transcription of the Speech

To examine the accuracy of the transcriptions of the statements, we used prepared recordings and reference data in the form of transcriptions of the spoken sentences. The research was carried out on 1,200 recordings, which included 30 statements for each of the two classes - statements with the intention to call for help and neutral statements, recorded twice by 10 people. The statements with the intention to call for help or containing descriptions of medical symptoms along with their transcription were taken from the KAGGLE *Medical Speech, Transcription, and Intent* dataset [19]. The neutral sentences were taken from the KAGGLE *TED Talks Transcripts for NLP* dataset containing transcripts of

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97573-8_2

talks given on various topics at TED scientific conferences [10]. The recordings were provided as input to the Vosk model, which produced their transcriptions. Transcriptions from the Vosk model were compared to reference transcripts to determine their accuracy. For this purpose, we used the following quality measures:

- *Word Error Rate* (WER) - the ratio of the total number of word errors (substitutions S, deletions D, or insertions I) to the number of words N in the reference transcription:

$$WER = \frac{S + D + I}{N}, \tag{5}$$

- *Sentence Error Rate* (SER) - the percentage of transcripts containing at least one incorrect word:

$$SER = \frac{F}{M}, \tag{6}$$

where S is the number of words replaced with other words relative to the reference transcription, D is the number of words omitted (deleted) relative to the reference transcription, I is the number of words inserted additionally to the reference transcription, N is the total number of words in the reference transcription, F is the number of transcriptions containing at least one incorrect word, M is the total number of transcriptions. Particular ratios of S/N, D/N, and I/N in our experiments are presented in Table 3.

When assessing the transcription quality, we did not consider capitalization and punctuation. Each transcription was changed to all lowercase letters, and punctuation was removed completely. Similarly to the wake-up word detection, experiments were performed on original recordings and recordings with added white noise. By default, the maximum amplitude of the noise was assumed to be equal to the maximum amplitude of the original recording. Additionally, we analyzed the recordings with additional noise, whose amplitudes were reduced by 20, 40, and 60 decibels compared to the default value.

Table 3: The ratio of the number of substitutions S, deletions D and insertions I to the number of all words in the reference transcript (μ - mean, σ - standard deviation)

	S		D		I	
	μ	σ	μ	σ	μ	σ
Original sound	0.063	0.107	0.016	0.044	0.012	0.046
With the white noise (-60 dB)	0.064	0.105	0.019	0.051	0.012	0.042
With the white noise (-40 dB)	0.079	0.129	0.094	0.244	0.010	0.036
With the white noise (-20 dB)	0.207	0.191	0.293	0.330	0.007	0.035
With the white noise (-0 dB)	0.119	0.077	0.879	0.077	0.000	0.000

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97573-8_2

For original sound, the speech-to-text transcription with the Vosk model achieved very good results (Table 4). The mean value of the WER error is 9.228%, and the standard deviation is 13.635%. Noteworthy is the median value of zero. This proves that a large group of recordings was transcribed flawlessly. Taking into account the specificity of the WER metric and its lowest possible value equal to zero, at least half of all recordings belong to this group.

Table 4: WER values for the speech-to-text transcription

	WER		
	Mean (%)	Standard deviation (%)	Median (%)
Original sound	9.228	13.635	0,000
With white noise (-60 dB)	9.520	13.236	0.000
With white noise (-40 dB)	18.368	27.102	9.100
With white noise (-20 dB)	50.737	33.952	50.000
With white noise (-0 dB)	99.907	0.737	100.000

As the amplitude of noise added to recordings changes, the ratios of substitutions, deletions, and insertions to the total words in the reference transcription also change. Higher noise amplitude increases deletions and substitutions (except at maximum amplitude) while reducing insertions (Table 3). Substitutions and deletions had the most significant impact on the WER metric. For original recordings and those with -60 dB noise, substitutions contributed about 0.06 to the ratio, with deletions three to four times lower. In contrast, for noise levels at -40, -20, and 0 dB, deletions had a greater influence. Substitution ratios ranged from 0.079 to 0.207, while deletion ratios spanned 0.094 to 0.879. Insertions had minimal impact on WER, regardless of noise level. The average insertion ratio was a maximum of 0.012, indicating the model rarely added unnecessary words to the transcription.

The SER error value increased with higher noise amplitude (Table 5). For original recordings, the SER was 7.917%, meaning 1,105 out of 1,200 recordings were transcribed correctly. Low-amplitude noise had little impact on results, with significant discrepancies only at noise levels of -40 dB or higher. Even with SER below 20%, these transcriptions can still aid in detecting intent to call for help when better-quality recordings are unavailable. However, the highest noise levels (-0 dB and -20 dB) produced the worst results, making transcriptions under these conditions unsuitable for emergency call procedures.

6.3 Detecting Intentions to Call for Help

For intention detection, we used the same transcription set as the previous experiment, dividing it 80:20 into balanced training and test sets with utterances expressing or not expressing the intent to call for help. Several models were trained and tested, including TF Simple, TF 1-layer Bi-LSTM, TF 2-layer

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97573-8_2

Table 5: SER values for the speech-to-text transcription ($M = 1200$)

	The number of transcripts containing at least one incorrect word F	SER (%)
Original sound	95	7.917
With white noise (-60 dB)	99	8.250
With white noise (-40 dB)	230	19.167
With white noise (-20 dB)	788	65.667
With white noise (-0 dB)	1200	100.000

Bi-LSTM, Embed&Class, Azure Cognitive Services, and GPT 3.5. The Embed&Class model uses three different kinds of input leading to three variations, standard rough sentences (Embed&Class), rough sentences with metadata like message titles, if available (Embed&Class-T), or messages titles and intention prompts, if available (Embed&Class-TP). The task was treated as a two-class classification, where models identified whether a statement indicated the intent to call for help. Test set utterances were classified by the models, and results were compared with reference data to build confusion matrices and calculate metrics, including accuracy, precision, sensitivity, and F1-score. We also measured the classification times for each model (see Table 6).

Table 6: Classification performance and time while detecting the intention to call for help (μ_T - mean execution time, σ_T - standard deviation)

	Accuracy	Precision	Sensitivity	F1-score	μ_T (s)	σ_T (s)
TF simple	0.900	0.910	0.887	0.899	0.038	0.012
TF 1-layer Bi-LSTM	0.887	0.869	0.912	0.890	0.044	0.003
TF 2-layer Bi-LSTM	0.894	0.889	0.900	0.894	0.042	0.005
Embed&Class	0,997	0,997	0.995	0,997	0.030	0.004
Embed&Class-T	0,998	0,998	0.997	0,998	0.031	0.004
Embed&Class-TP	1.000	1.000	1.000	1.000	0.032	0.004
Azure Cognitive Services	0.981	0.975	0.988	0.981	0.207	0.038
GPT 3.5	0.881	0.867	0.900	0.883	0.795	0.219

The accuracy of local TensorFlow models working on the edge device did not exceed 90%. Among them, the *TF simple* model had the highest precision, making it the most effective at identifying emergency calls, though it detected about 1 in eleven calls incorrectly. The *1-layer Bi-LSTM* model achieved the highest sensitivity, but cases of missed emergency calls remain. Differences in F1-scores among these models were minimal. GPT 3.5 performed the worst among remote models, with an accuracy of 88.1% and the lowest precision and F1-score. In contrast, Azure Cognitive Services delivered better results, with a 98.1% accuracy. However, the best results are obtained by the local Embed&Class models that outperforms all models, even remote prompt-based ones, with a minimum

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97573-8_2

of 99.7% accuracy. Although simpler than the local TensorFlow models, these models even reaches 100% accuracy whenever additional metadata are used for learning (Embed&Class-TP).

Inference times for statement classification were shortest with TensorFlow and Embed&Class models running locally on the edge device, averaging 30–44 milliseconds, enabling near-instant detection of help intent. Remotely invoked models were slower due to communication delays. GPT 3.5 had the worst performance, averaging nearly 800 milliseconds, with some cases taking up to 2,334 milliseconds—about 20 times slower than local models. Azure Cognitive Services performed better, with an average classification time of 207 milliseconds, a reasonable result for a cloud-based solution requiring internet access. While both remote services are acceptable for this application, their differing speeds may affect user experience.

7 Discussion and Concluding Remarks

Our experiments demonstrated that we can successfully perform automatic help calls through speech and text data analysis at the Edge. An interesting finding from these experiments is that lightweight AI/ML models running on Edge devices may outperform remote prompt-based models for this specific task, whilst working locally on smartphones. This not only accelerates the initiation of help requests or caregiver alerts while challenging previous results, such as those presented in [20], which suggested that cloud-based services often offer better accuracy, but also highlights how these services can be hindered by limited or delayed internet connectivity.

The rapid evolution of language models, including their lighter versions, is significantly transforming this landscape, notably enhancing the efficiency and accuracy of local approaches. These advancements ensure continuous, unobtrusive, and discreet monitoring of seniors, enhancing their sense of security without infringing on their independence. In fact, local solutions may even strengthen their autonomy by reducing the reliance on remote systems. Regardless of the approach used—whether local Edge-based, remote cloud-based, or hybrid—all of them contribute to enhancing the lives of older adults by making smart technologies accessible across generations and promoting their inclusion in modern society.

Acknowledgments

This paper was supported by the Reactive Too project that has received funding from the European Union's Horizon 2020 Research, Innovation and Staff Exchange Programme under the Marie Skłodowska-Curie Action (Grant Agreement No871163), the pro-quality grant (02/100/RGJ25/0041) of the Rector of the Silesian University of Technology (SUT), Gliwice, Poland, Statutory Research funds of the Department of Applied Informatics at SUT (grants No 02/100/BK_25/0044). Scientific work published as part of an international project

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97573-8_2

co-financed by the program of the Minister of Science and Higher Education entitled "PMW" in the years 2021 - 2025; contract no. 5169/H2020/2020/2.

References

1. Global health and aging. Tech. Rep. 7737, World Health Organization (2011)
2. Aakesh, U., Rajasekaran, Y., Bethanney Janney, J.: Wristband for elderly individuals: Esp-32 and arduino nano enabled solution for health monitoring and tracking. In: 2023 IEEE Asia-Pacific Conference on Geoscience, Electronics and Remote Sensing Technology (AGERS). pp. 26–33 (2023)
3. Albert M. Cook, Janice Miller Polgar, P.E.: Assistive Technologies. Principles and Practice. Elsevier, St. Louis, Missouri (2016)
4. AlphaCephei: Vosk speech recognition (2024), https://alphacephei.com/vosk/
5. AlphaCephei: Vosk speech recognition models (2024), https://alphacephei.com/vosk/models/
6. 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)
7. Banos, O., Villalonga, C., Damas, M., Gloesekoetter, P., Pomares, H., Rojas, I.: Physiodroid: Combining wearable health sensors and mobile devices for a ubiquitous, continuous, and personal monitoring. The Scientific World Journal **2014**, 1–11 (09 2014). https://doi.org/10.1155/2014/490824
8. 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)
9. 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)
10. Corral Jr., M.: TED talks transcripts for NLP (2020), https://www.kaggle.com/datasets/miguelcorraljr/ted-ultimate-dataset
11. 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)
12. Hannun, A.Y., Case, C., Casper, J., Catanzaro, B., Diamos, G.F., Elsen, E., Prenger, R.J., Satheesh, S., Sengupta, S., Coates, A., Ng, A.: Deep speech: Scaling up end-to-end speech recognition. ArXiv **abs/1412.5567** (2014), https://api.semanticscholar.org/CorpusID:16979536
13. Hosseinzadeh, M., Koohpayehzadeh, J., Ghafour, M.Y., Ahmed, A.M., Asghari, P., Souri, A., Pourasghari, H., Rezapour, A.: An elderly health monitoring system based on biological and behavioral indicators in internet of things. Journal of Ambient Intelligence and Humanized Computing **14**, 5085–5095 (05 2023)
14. Hou, L., Latif, J., Mehryar, P., Zulfiqur, A., Withers, S., Plastropoulos, A.: Iot based smart wheelchair for elderly healthcare monitoring. In: 2021 IEEE 6th International Conference on Computer and Communication Systems (ICCCS). pp. 917–921 (2021)
15. Huggins-Daines, D., Kumar, M., Chan, A., Black, A., Ravishankar, M., Rudnicky, A.: Pocketsphinx: A free, real-time continuous speech recognition system for hand-held devices. In: 2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings. vol. 1, pp. I–I (2006)

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97573-8_2

16. Jain, R., Semwal, V.B.: A novel feature extraction method for preimpact fall detection system using deep learning and wearable sensors. IEEE Sensors Journal **22**(23), 22943–22951 (2022)
17. Lersilp, S., Putthinoi, S., Lerttrakarnnon, P., Silsupadol, P.: Development and usability testing of an emergency alert device for elderly people and people with disabilities. The Scientific World Journal **2020**, 1–7 (02 2020)
18. Md Sabri, S., Annuar, N., Abdull Rahman, N.L., Musairah, S., Abd Mutalib, H., Subagja, I.: Major trends in ageing population research: A bibliometric analysis from 2001 to 2021. Proceedings **82**, 19 (2022)
19. Mooney, P.: Medical speech, transcription, and intent (2019), `https://www.kaggle.com/datasets/paultimothymooney/medical-speech-transcription-and-intent`
20. Mrozek, D., Kwaśnicki, S., Sunderam, V., Małysiak-Mrozek, B., Tokarz, K., Kozielski, S.: Comparison of speech recognition and natural language understanding frameworks for detection of dangers with smart wearables. In: Paszynski, M., Kranzlmüller, D., Krzhizhanovskaya, V.V., Dongarra, J.J., Sloot, P.M. (eds.) Computational Science – ICCS 2021. pp. 471–484. Springer International Publishing, Cham (2021)
21. Picovoice: Porcupine wake word (2024), `https://picovoice.ai/platform/porcupine/`
22. Povey, D., Ghoshal, A., Boulianne, G., Burget, L., Glembek, O., Goel, N., Hannemann, M., Motlíček, P., Qian, Y., Schwarz, P., Silovský, J., Stemmer, G., Vesel, K.: The kaldi speech recognition toolkit. IEEE 2011 Workshop on Automatic Speech Recognition and Understanding (01 2011)
23. Qian, Z., Lin, Y., Jing, W., Ma, Z., Liu, H., Yin, R., Li, Z., Bi, Z., Zhang, W.: Development of a real-time wearable fall detection system in the context of internet of things. IEEE Internet of Things Journal **9**(21), 21999–22007 (2022)
24. Radford, A., Kim, J.W., Xu, T., Brockman, G., McLeavey, C., Sutskever, I.: Robust speech recognition via large-scale weak supervision. In: Proceedings of the 40th International Conference on Machine Learning. ICML'23, JMLR.org (2023)
25. Trabelsi, A., Warichet, S., Aajaoun, Y., Soussilane, S.: Evaluation of the efficiency of state-of-the-art speech recognition engines. Procedia Computer Science **207**, 2242–2252 (2022), knowledge-Based and Intelligent Information Engineering Systems: Proceedings of the 26th International Conference KES2022
26. United Nations, Department of Economic and Social Affairs, Population Division: World Population Ageing 2019. United Nations, New York (2020)
27. United Nations, Department of Economic and Social Affairs, Population Division: World Population Ageing 2020 Highlights: Living arrangements of older persons. United Nations, New York (2020)
28. United Nations, Department of Economic and Social Affairs, Population Division: World Social Report 2023: Leaving No One Behind In An Ageing World. United Nations, New York (2023)
29. Vinotha, R., Hepsiba, D., Vijay Anand, L., D.: Leveraging openai whisper model to improve speech recognition for dysarthric individuals. In: 2024 Asia Pacific Conference on Innovation in Technology (APCIT). pp. 1–5 (2024)
30. Zhang, Q., Ren, L., Shi, W.: Honey: a multimodality fall detection and telecare system. Telemedicine journal and e-health: the official journal of the American Telemedicine Association **19 5**, 415–29 (03 2013)

ICCS Camera Ready Version 2025
To cite this paper please use the final published version:
DOI: 10.1007/978-3-031-97573-8_2