applied sciences

MDPI

Article

ODIN112–AI-Assisted Emergency Services in Romania

Dan Ungureanu [1], Stefan-Adrian Toma [2], Ion-Dorinel Filip [1], Bogdan-Costel Mocanu [1], Iulian Aciobăniței [2], Bogdan Marghescu [2], Titus Balan [3,4], Mihai Dascalu [1,5,*], Ion Bica [2] and Florin Pop [1,6]

1 Computer Science & Engineering Department, University Politehnica of Bucharest, 060042 Bucharest, Romania
2 Military Technical Academy Ferdinand I, 050141 Bucharest, Romania
3 Department of Electronics and Computers, Transilvania University of Brasov, 500036 Brasov, Romania
4 ATOS Convergence Creators, 500090 Brasov, Romania
5 Academy of Romanian Scientists, Str. Ilfov, Nr. 3, 050044 Bucharest, Romania
6 National Institute for Research & Development in Informatics-ICI Bucharest, 011555 Bucharest, Romania
* Correspondence: mihai.dascalu@upb.ro

Abstract: The evolution of Natural Language Processing technologies transformed them into viable choices for various accessibility features and for facilitating interactions between humans and computers. A subset of them consists of speech processing systems, such as Automatic Speech Recognition, which became more accurate and more popular as a result. In this article, we introduce an architecture built around various speech processing systems to enhance Romanian emergency services. Our system is designed to help the operator evaluate various situations with the end goal of reducing the response times of emergency services. We also release the largest high-quality speech dataset of more than 150 h for Romanian. Our architecture includes an Automatic Speech Recognition model to transcribe calls automatically and augment the operator's notes, as well as a Speech Recognition model to classify the caller's emotions. We achieve state-of-the-art results on both tasks, while our demonstrator is designed to be integrated with the Romanian emergency system.

Keywords: Automatic Speech Recognition; Speech Emotion Recognition; Romanian language; emergency services

Citation: Ungureanu, D.; Toma, S.-A.; Filip, I.D.; Mocanu, B.-C.; Aciobăniței, C.; Marghescu, B.; Balan, T.; Dascalu, M.; Bica, I.; Pop, F. ODIN112–AI-Assisted Emergency Services in Romania. *Appl. Sci.* **2023**, *13*, 639. https://doi.org/10.3390/app13010639

Academic Editors: Kyungyong Chung and Ellen J. Hong

Received: 2 December 2022
Revised: 24 December 2022
Accepted: 27 December 2022
Published: 3 January 2023

1. Introduction

In recent years, the developments in Artificial Intelligence (AI) and, more specifically, Natural Language Processing (NLP) made it a popular choice for various tools we use every day, such as automatic translation, grammar correction, accessibility tools, and chatbots. NLP technologies facilitate the interaction between humans and computers by allowing machines to transcribe and understand spoken language.

Automatic Speech Recognition (ASR), also known as speech-to-text, is the study of computational linguistics that focuses on processing speech signals with the objective of transcribing spoken language into text. This process is usually effortless for real humans, but it is a challenge for computers due to the various complexities of oral communication.

The first things that can be observed are the different accents or dialects, as well as changes in pitch, tempo, or volume. Moreover, transcribing and understanding speech becomes even more difficult as we consider the cognitive processes required for observation and understanding the context in which the words are transmitted. The context cannot be transformed into inputs for artificial neural networks as they are heavily influenced by life experiences, together with the personal perception of the surrounding world. Finally, challenges are present at a lower level concerning communication (e.g., background noise, compressed audio) or at a higher level (e.g., punctuation, capitalization, or text formatting in general).

Emergency services are present worldwide and, in Romania, the service is available to the general population by calling the quick phone number "112", similar to "911" in

the United States of America or "999" in the United Kingdom. The Romanian emergency system follows the European Telecommunications Standards Institute (ETSI) standardization direction of the NG112 (Next Generation 112) specification corresponding to the European emergency services. NG112 brings new functionalities, especially at the level of multimedia communications, which opens the opportunity to implement new types of services. Speech-recognition-based intelligence can ease the work of emergency services dispatchers and provide relevant information for critical cases.

Though there are not many AI-enhanced emergency systems in operation, such tools could help analyze calls and messages and support faster decisions. In return, AI tools could reduce waiting times and save more lives by assisting operators in several use cases like filtering abusive calls more efficiently and providing contextualized suggestions to operators in specific situations. An example is the Danish company Corti [1], which developed an AI system to analyze emergency calls and predict cardiac arrests more accurately outside the hospital. According to Madsen et al. [2], emergency dispatchers fail to identify about 25% of cases of cardiac arrest and therefore lose the possibility of giving instructions for resuscitation. Artificial Intelligence could detect a cardiac arrest also based on historical data given a predisposition of the caller for such conditions. The developed algorithm listens to the call in real time and alerts the operator if the caller has a cardiac arrest. In 2018, EENA launched a pilot project with Corti to test how this system could work; the project was carried out in various locations in France and Italy. In another implementation, ML2Grow (https://ml2grow.com/case/ai-system-ensures-faster-handling-of-emergency-calls/, accessed on 1 December 2022) was used to identify accidental calls. Based on the first seconds of a call, the "Hazira Digital" system determined whether the call was accidental or intentional (real).

This article presents our solution for the Romanian emergency systems with a focus on the underlying deep models and the overall architecture of the developed system, our methods for data acquisition, and the training of models for a low-resource language.

Our main contributions are as follows:

- To our knowledge, the largest high-quality speech dataset of more than 150 h for Romanian available at https://echo.readerbench.com/ (accessed on 1 December 2022);
- An architecture consisting of two central components, namely Automatic Speech Recognition and Emotion Recognition models that achieve state-of-the-art results, integrated with the Romanian emergency system and designed to support the operators take timely decisions.

The paper is structured as follows. After the introduction, the second section presents related work, while the third section details the proposed system architecture, the speech and emotion recognition algorithms, as well as the corpora used for training the models. The fourth section unveils the results for speech and emotion recognition. The results are discussed in the fifth section, while the paper ends with conclusions.

2. Related Work

In this section, we present state-of-the-art methods for Automatic Speech Recognition and Speech Emotion Recognition.

2.1. Automatic Speech Recognition

Automatic Speech Recognition has been a subject of study for a long time. However, significant improvements to make it viable for text transcription were made only in the last decade. The underlying process is made up of several steps, the most common and important ones being feature extraction, phoneme detection, word composition, and text transcription.

The conventional implementation used to be a hard-coded pipeline and acoustic model tweaked manually until the accuracy was good enough for the specific domain it was used in. As the pipeline was built on top of different modules that used various algorithms for a particular task, optimizing the system implied adjusting each component.

The optimization process was laborious and sometimes reached the local optimum for each component instead of the global optimum [3].

The newer methods leverage end-to-end learning, training a single Deep Neural Network (DNN) model with multiple layers that replace the traditional pipelines. Neural networks can process inputs and outputs of arbitrary length and take advantage of local spatial coherence extracting more useful information with a relatively low computational cost. These are just two of the characteristics that make them ideal for audio and speech analysis. Furthermore, these systems can be classified into regular DNN-based models or hybrid models, which leverage Deep Neural Networks and any other set of technologies, such as the long-established Hidden Markov Models.

2.1.1. Romanian Datasets

Even though there is a decent amount of speech recordings available for the Romanian language, and this quantity continues to increase steadily, it remains a low-resource language by modern deep learning standards. Moreover, part of this data has low quality or is too noisy for training or evaluation. Table 1 presents all available speech datasets available for the Romanian language, summing up to around 300 h of recordings.

Table 1. Available speech datasets for the Romanian language.

Data Set	Duration	Recordings	Unique Transcripts	Unique Word Count	Speakers
Chamber of Deputies (eval)	4 h	296	296	7084	N/A
IIT	18 h	8877	8619	30,327	98
RACAI	11 h	3404	2646	15,877	3
RADOR	50 h	16,180	15,530	50,782	N/A
RASC	4 h	2976	2972	14,113	N/A
Robin	1 h	400	194	2165	4
RoDigits	37 h	15,389	100	10	154
Romanian Read-Speech [4]	99 h	136,120	11,924	18,485	164
SSC (eval)	5 h	3135	3008	11,291	N/A
SWARA	21 h	19,292	1803	6102	17
Various	32 h	18,568	15,419	36,174	3444+

The "various" dataset represents miscellaneous recordings that were obtained from unknown sources that are too small in size to be representative on their own but have been collected into a single, larger dataset that can be used for training. This is characterized by a large number of speakers present in the dataset.

2.1.2. Hidden Markov Model-based Architectures

Hidden Markov Models (HMMs; Rabiner and Juang [5]) are a statistical Markov model in which the states are not observable. HMMs learn about the initial Markov process by observing a different process whose behavior depends on the modeled system. Gaussian Mixture Models (GMMs; Reynolds [6]) are a probabilistic model that assumes all data points are generated from a mixture of a finite number of Gaussian distributions with unknown parameters.

In speech recognition, HMMs are used together with GMMs. The purpose of the Hidden Markov Model is to consider the temporal dependencies between the acoustic features which represent the context. In contrast, the Gaussian Mixture Model is used to classify each frame, which represents the uttered phonemes. Paired with a pronunciation lexicon and a language model, HMM-GMM models (Figure 1) can transform a given sound wave into a word sequence. Each audio frame of the input wave is processed while acoustic features (for example, Mel-Frequency Cepstral Coefficients or MFCC for short) are extracted. The sentences are split into words, and then each word is converted into a

sequence of phonemes using a phonetic dictionary or an additional grapheme-to-phoneme model to determine the phonemes automatically.

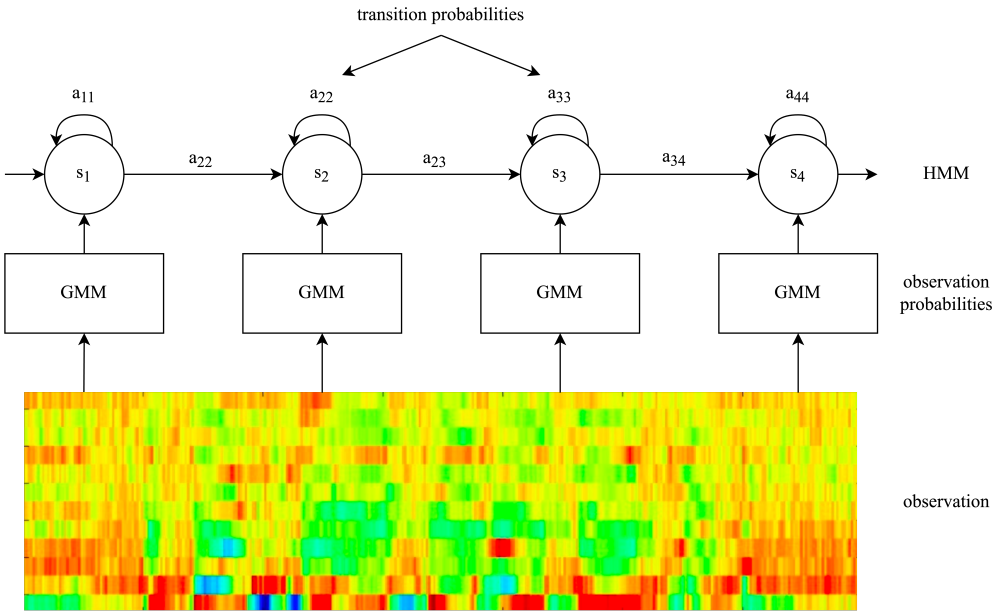


Figure 1. Hidden Markov Model-based model for Automatic Speech Recognition.

CMUSphinx [7] is the most well-known speech-to-text solution, a speech recognition development toolkit heavily used for experimenting with HMM-GMM-based speech recognition architectures (Figure 2). It supports multiple types of acoustic models: continuous, semicontinuous, and phonetic-tied models. The best results obtained with CMUSphinx for the English language were obtained after training with LibriSpeech (100 h of clean audio) (https://abuccts.blogspot.com/2017/08/gsoc-2017-with-cmusphinx-post-11.html, accessed on 1 December 2022): 19.4% WER (continuous model) and 30.3% WER (phonetically tied model). The best results for the Romanian language vary between 3.8% WER and 23.17% WER, the former result being obtained [8] for a subset of a small and clean dataset, while the latter for a much more difficult evaluation dataset.

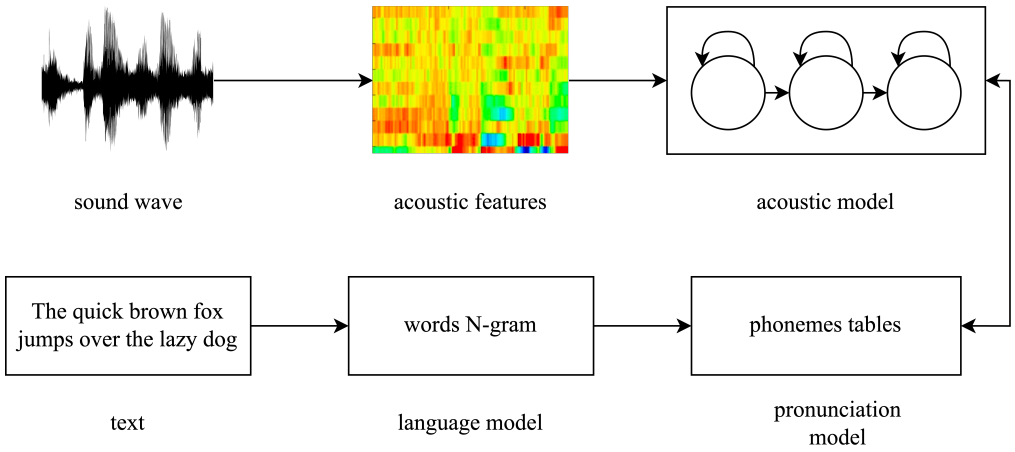


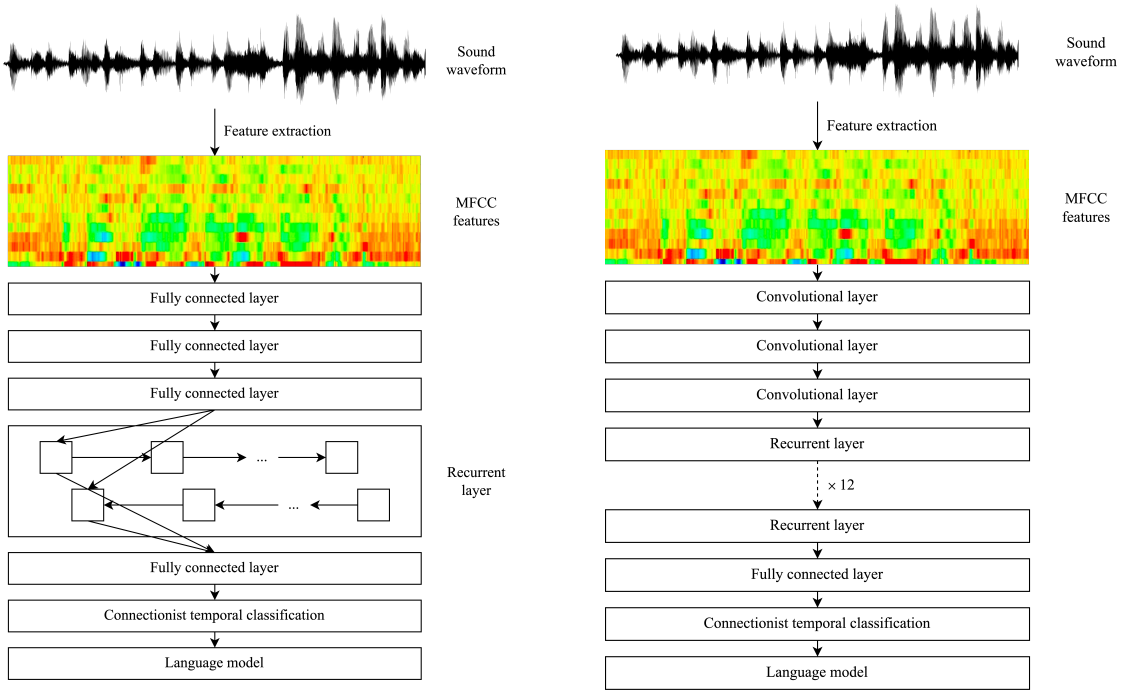
Figure 2. Architecture of Hidden Markov Model-based Automatic Speech Recognition system.

2.1.3. Deep Neural Network-Based Architectures

In the field of Automatic Speech Recognition, a breakthrough consisted of the use of Deep Neural Networks (DNNs) for developing end-to-end speech-to-text systems. They successfully replaced the older and more complex traditional systems with a single neural net-

work. DNNs have a simpler implementation and consistently outperform previous solutions. The variety of networks used ranges from simple Recurrent Neural Networks (RNNs) using Long Short-Term Memory (LSTM; Hochreiter and Schmidhuber [9]) to Transformer models using Attention Mechanisms [10], mixing supervised and unsupervised learning, solving both tasks of speech recognition and language modeling.

Deep Speech [3] (see Figure 3a) is one of the most known solutions built using a Recurrent Neural Network (a single recurrent layer), which obtained a 16% WER after training the model using the Switchboard dataset composed of 300 h of labeled speech data. The second version, DeepSpeech 2 [11] (see Figure 3b), uses more recurrent layers and further reduced the WER to 12.73% when trained on LibriSpeech. The most accurate model built using deep neural networks is Whisper [12], which achieved a WER of 2.7% on the LibriSpeech test-clean dataset. Whisper almost halved the overall WER down to 12.9% in comparison to wav2vec 2.0 [13], another renowned Automatic Speech Recognition model, which achieved a WER of 29.5%.



(**a**) DeepSpeech 1 (**b**) DeepSpeech 2

Figure 3. Architecture of DeepSpeech Automatic Speech Recognition models.

2.1.4. Hybrid Architectures

The common perception is that the two solutions (i.e., Hidden Markov Model-based and Deep Neural Network-based architectures) are competitors in the field of Automatic Speech Recognition. Still, experiments indicate that the two technologies can be used together to benefit from the advantages of both types of models. In the past, they have shown remarkable accuracy and flexibility. Deep Neural Networks can be used for providing probability estimates for the state of the HMM, or smaller Hidden Markov Models can be used for aligning training data for the DNN.

One well-known open-source toolkit for speech recognition that employs a hybrid architecture is Kaldi [14]. The best-performing recipe provided by Kaldi obtained a WER of 8.76% when trained on the LibriSpeech dataset. The best result for a hybrid architecture, combining HMMs and DNNs, and the same dataset is a 5.0% WER.

2.1.5. Transformer Architectures

Transformers [10] have changed the NLP landscape in terms of employed models. Compared to RNN and convolutional networks, Transformers can be trained significantly faster and, for many use-cases, achieved state-of-the-art performances [10].

A powerful framework for transformers ASR models is wav2vec 2.0 [13]. The main advantage of the framework is that models can be pretrained on large unlabeled data and then fine-tuned on small labeled datasets. The authors achieved state-of-the-art results (WER 4.8/8.2 on test-clean/other of LibriSpeech) with only 10 min of labeled data. Even though wav2vec 2.0 for ASR is capable of producing good results, it can be improved using a language model. Wav2vec 2.0 offers support for the n-gram language model based on KenLM [15].

2.2. Speech Emotion Recognition

Speech Emotion Recognition is the task of identifying the emotion of the speaker from audio signals. Human speech is much more than just words, and humans can efficiently perform it as a natural part of our day-to-day oral communication, observe the other acoustic properties (pitch, tempo, volume, etc.) and parse it accordingly (add or modify information). From a machine learning perspective, this task is a classification problem, like sentiment analysis, but it runs on raw speech instead of transcribed text. Emotion classification is not an easy task. There are two models used for emotion classification: the discrete model and the dimensional model. The discrete emotional model is based on the six basic innate emotions described by Eckman et al. [16]: sadness, happiness, fear, anger, disgust, and surprise. All other emotions are obtained by combining the basic ones. The dimensional model uses a small number of characteristics, such as valence (wheater positive or negative) and activation (i.e., intensity). Though simpler, the dimensional model is unintuitive for most people and requires special training for emotional corpora labeling. Hence, most emotional speech classifiers use the discrete model.

2.2.1. Corpora

Since speech emotion recognition is basically a classification problem, emotional speech corpora play an integral part in the performance of such a system. Nevertheless, obtaining such corpora is difficult. Unlike speech recognition, participants have to either mimic an emotion or record an actual one and manually annotate each entry. Hence, there are three types of datasets: simulated, induced, and natural. Simulated emotional speech corpora such as IEMOCAP [17], EMODB [18], Toronto Emotional Speech Set [19], Danish Emotional Speech (DES) [20], and Italian Emotional Speech Database (EMOVO) [21] are created by actors (professional or amateur). Induced ones, such as eNTERFACE'05 Audio-Visual Emotion Database [22] are recorded while emotion is triggered by some external stimulus. Natural emotions corpora (e.g., AFEW-VA Database [23]) are extracted from real-life recordings.

Our choice for the ODIN112 system is the only available Romanian corpus, EmoIIT [24]. We also trained and tested the developed classifier using the EmoDB [18] database to compare our implemented classifier with other published results.

EmoIIT [24] is an emotional speech corpus recorded in Romanian. It is similar in structure to EmoDB, and, to our knowledge, the only Romanian emotional speech corpus available. The speakers are amateur actors, ages 20 to 22 years old, mostly students. The dataset contains 523 recordings, split between 7 emotions: anger, boredom, fear, happiness, sadness, disgust, and neutral.

The Berlin Emotional Speech (EmoDB) [18] dataset is an emotional speech corpus created by the Institute of Communication Sciences, Technical University, Berlin, Germany. It was recorded by ten professional actors-five men and five women, in German. The EmoDB dataset contains a total of 535 records representing 7 emotions: anger, boredom, fear, happiness, sadness, disgust, and neutral. Recordings were made at a sampling rate of 48 kHz and then down-sampled to 16 kHz.

2.2.2. Models

Almost all emotional speech classification algorithms use four classes of features: prosodic, spectral, voice quality features, and Teager energy operator [25]. Currently, most emotional speech classifiers are based on deep neural networks or a combination of traditional classifiers and deep neural networks, such as in [26–31]. Results vary between 50% to more than 92% in overall accuracy [25]. A big factor in the overall performance is the corpus used for training and testing.

3. Method

3.1. Corpus

One advantage of using deep neural networks is that they can learn from much larger datasets. Many modern systems have been developed thanks not only to the advances in neural architectures but also because of the availability of data, storage capabilities, and the increased processing power of new computer hardware.

The largest limitation in training accurate speech processing models for the Romanian language is the lack of data. Romanian is far from being a popular language and it is spoken by approximately 23 million people. The total duration of publicly available resources is approximately 300 h, out of which just about half are high quality and can be used for producing relevant models. As a comparison, much more popular languages such as English or Chinese have over 1000 h of labeled speech data and considerably more unlabeled speech datasets. A recent neural network model named Whisper [12] has been trained using 438,218 h of English hours and over 117,113 h, among the most popular languages being Chinese (23,446 h), German (13,344 h), Spanish (11,100 h), Russian (9761 h), French (9752 h), Portuguese (8573 h), Korean (7993 h), and Japanese (7054 h).

Data Acquisition

Datasets are one of the most important aspects of machine learning, and currently, audio datasets of hundreds of hours of labeled speech are usual. New models have been trained with over hundreds of thousands of hours of recordings. One remaining problem consists of finding or collecting high-quality training data (clear recordings and correct transcripts). In general, high-quality training datasets must be manually curated and are difficult and expensive to produce because of the amount of necessary time involved in recording and labeling data.

New methods that make use of unsupervised learning or semisupervised learning can also include unlabeled entries. Still, in these cases, the data quantity requirements increase significantly, and data acquisition remains a tiresome task. Another idea is to use existent labeled data for bootstrapping a model that can transcribe the unlabeled data. Nevertheless, this method is not ideal either because it is subject to recognition errors.

At this moment, there are several initiatives to crowdsource a dataset for the Romanian language. Two of the most successful examples are Mozilla Common Voice (https://commonvoice.mozilla.org, accessed on 1 December 2022) (40 h) and Echo (https://echo.readerbench.com, accessed on 1 December 2022) (our project with more than 150 h, see Table 2).

The data was recorded with the help of more than 200 volunteers, most of whom were students at the University Politehnica of Bucharest aged between 20 and 24 years. Tables 3 and 4 depict the distributions of their ages and gender; however, only about half of the volunteers disclosed this information.

Data augmentation techniques are a common solution when the data is sparse. One such solution is SpecAugment [32], one of the most popular and efficient augmentation methods for speech. SpecAugment processes audio data by altering the audio spectrogram instead of the raw input audio waveform. This method is simple as it consists only of a set of basic operations: warping, masking of blocks in the frequency domain, or masking of time steps).

Table 2. Subsets of the collected speech corpora for the Romanian language.

Data Set	Duration	Recordings	Unique Transcripts	Unique Word Count	Speakers
Drama	6 h	4146	523	2724	77
Emergencies	10 h	6937	1309	822	205
Legal	13 h	4101	577	3079	76
Narratives	23 h	10,978	482	7961	81
News	11 h	4879	168	4391	76
Poetry	3 h	917	112	1836	71
Wikipedia	92 h	35,754	1001	7509	212
Echo Total	155 h	66,795	4059	22,521	215

Table 3. Statistics about the age of the volunteers.

Age	19	20	21	22	23	24	25+	N/A	Total
Count	3	15	10	37	22	14	8	106	215

Table 4. Statistics about the gender of the volunteers.

Gender	Male	Female	N/A	Total
	41	71	103	215

3.2. ODIN112 Architecture

Our ODIN112 architecture (see Figure 4) is built around various speech processing systems to enhance Romanian emergency services. It includes an Automatic Speech Recognition module to transcribe calls automatically and augment the operator's notes, as well as a module to classify the caller's emotions. ODIN112 is meant to support the operator in evaluating the situation with the end goal of reducing the response times of emergency services (e.g., ambulances, firefighters, and police).

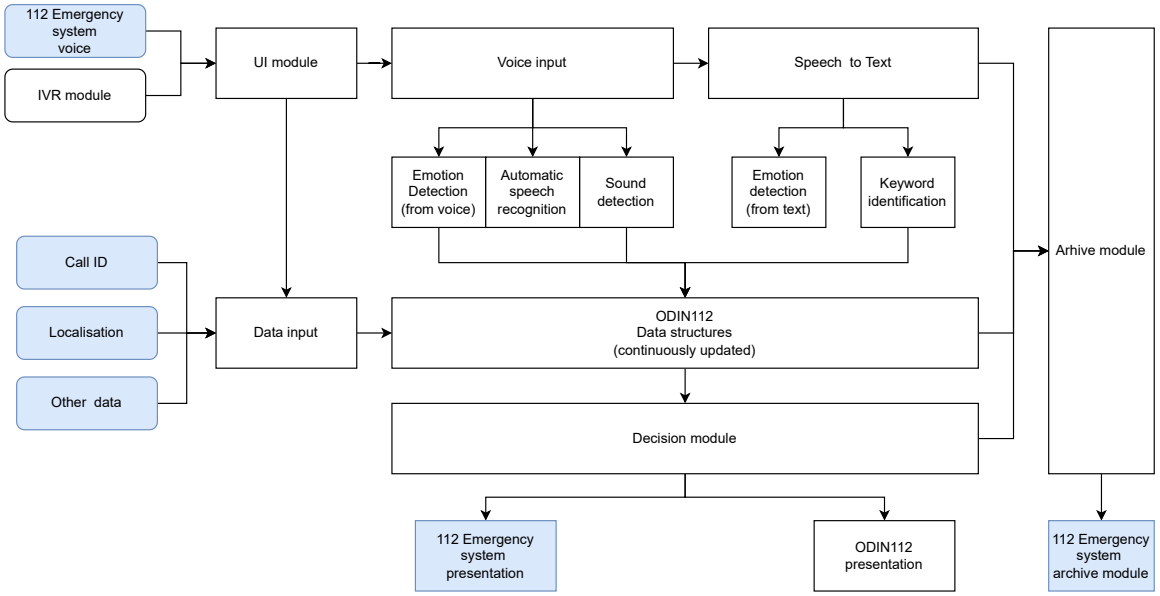


Figure 4. ODIN112 Architecture: Functional Components (the existing emergency services system components are represented in blue).

The general architecture of the information system is composed of a telephone exchange proxy, the IVR module (presented in detail in [33]), the ODIN112 user interface, and

a set of microservices that provide the services of the scene and acoustic event classification, emotion analysis, speech transcription. The Manager component uses the microservices to handle the real-time processing of the audio streams associated with the inbound calls. All modules are delivered as standard (Docker-like) containers, and the communication between them is done through simple or standard interfaces. Those decisions make it easy to be integrated into SNUAU, the Unique National Emergency Call System.

SNUAU is composed of single emergency call centers, integrated emergency dispatch centers, and emergency dispatch centers of specialized intervention agencies. Both the single centers for emergency calls, operated by the Romanian Special Telecommunication Services (STS), and the emergency dispatches of specialized intervention agencies are organized at the level of each county in Romania.

STS and their partners developed the current infrastructure of the emergency services in the direction of the NG112 specifications. Our demonstrator is not fully integrated with the new Romanian emergency system which is not yet in full operation, yet we considered the usage of standard and open interfaces for rapid further integration: the data format for intermodule communication and archiving is done in a JSON format to facilitate further integration into the existing emergency system; archiving is done in a system that enables rapid scaling—i.e., HDFS (Hadoop File System); our decision support system based on keywords can be easily triggered via REST services; the decision support system identifies the emergency index nodes that are usually used by the existing emergency system to classify the emergency cases based on type into different subcategories. Thus, the response to the incidents is triggered (e.g., the decision to involve an ambulance). In order to achieve the online processing of the audio streams, we used TCP sockets to continuously feed the processing modules with new parts of the audio call. The requirements are different compared to real-time communication (such as VoIP), which is usually done using UDP/RTP since the automated process requires reliable communication.

Our solution was developed to enable multiple parallel calls between the operator, the caller, and the agencies providing the emergency services. The communication was optimized for real-time processing, different modules can benefit from the entire call without having to maintain a state between requests (new call samples are added to the TCP flow dedicated for that specific call), and the modules have the possibility to return results at times independent of receiving samples. Based on our experimental evaluation, the overall experience is similar to observing automated captioning of live streaming; intermediary results are published into the operator's interface, and the automated feedback from our solution is provided to the operator without delays.

The integration between the components inside the ODIN112 pilot is depicted in Figure 5. When a new call arrives, the ODIN112 Manager component notifies the backend of the user interface module to register the call. The ODIN112 Manager also starts multiplying the audio stream to the transcription, noise identification, and voice sentiment analysis services. The communication between the manager and the processing services/modules is done using bidirectional TCP streams, allowing new audio fragments to be delivered for processing at any time (achieving real-time processing) and each service module to provide new results as soon as they are available.

Data-wise communication (consisting of Transcriptions, Emotion, and Sound detection results) between the manager and the ODIN112 backend application is made through Kafka queues ensuring scalable, fault-tolerant stream processing. Using such an implementation helps decouple the producers and the consumers of the system while still providing reliable processing for each provided message. Each call is handled on a Kafka topic defined as the unique ID of the call. When the manager observes the end of a call, it notifies all the other components (including the backend module), waits for the last results from each of the processing services, and then decommissions the streams for the call that just ended. Kafka offers a fault-tolerant implementation for ordered queues of messages. When compared to using another open-source implementation of a consensus algorithm (such as Raft),

Kafka comes with the advantage of being highly configurable in respect of replication and fault tolerance.

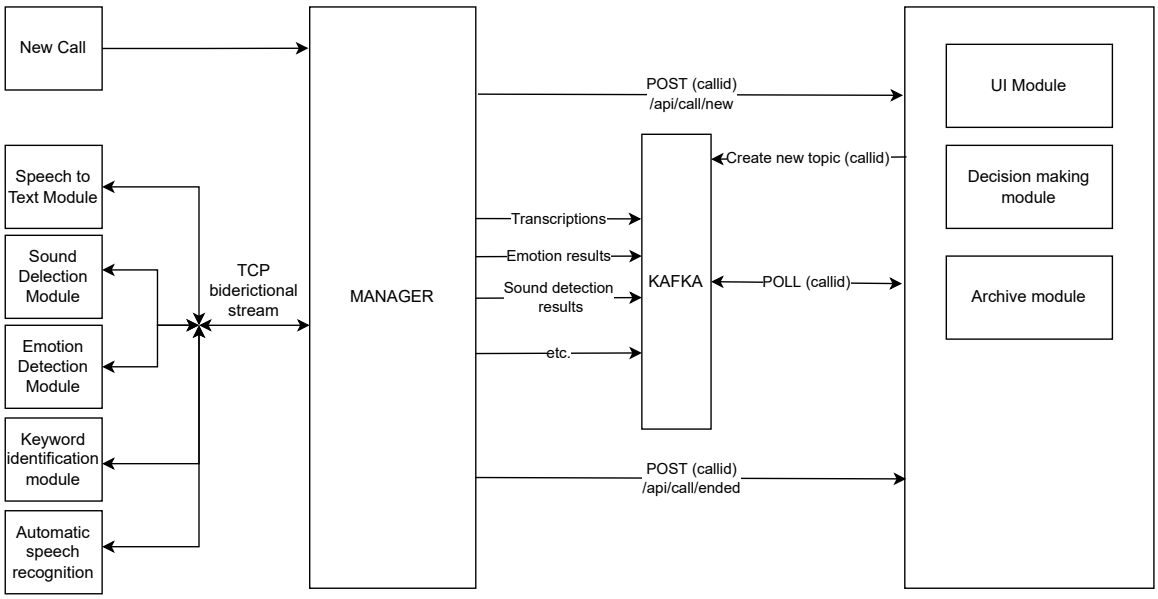


Figure 5. ODIN112 Architecture: Communication and integration of components.

The text obtained from the transcript is sent to the keyword identification and text sentiment analysis services. Keyword identification considers predefined word lists of interest, whereas the text analysis service for emotion detection from text considers a pretrained Romanian BERT model [34]. All obtained results are transmitted to the decision module and recommendations will be triggered. As soon as the results start to be generated, they are sent to the presentation interface and to the archiving module for long-term storage. The archive module is backed by a Hadoop Distributed File System (HDFS) to provide scalable, highly fault-tolerant storage required for the monitorization of any critical system. These later components are not included in the scope of this paper.

3.3. Automatic Speech Recognition Model

For this project, we used the Kaldi toolkit to develop our model (see Figure 6) and designed a framework for ingesting our training data. As with any Kaldi recipe, the processing logic happens in several stages, each representing a step of the data preparation process or training a model. The most important steps are extraction of the acoustic features from input recordings, training several HMM-GMM models that become gradually more complex and use more data, and finally training a Deep Neural Network model that uses twelve Time Delay Neural Network Factorization (TDNNF) layers. Time Delay Neural Networks classify patterns with shift-invariance and learn context at each layer of the network. This network does not require explicit segmentation prior to classification.

The features extracted for the acoustic features are the Mel Frequency Cepstral Coefficients (MFCC; Mermelstein [35]), followed by applying Cepstral Mean and Variance Normalization (CMVN). The MFCCs are extracted from frames of 25 ms with a stride of 10ms. For the input of the TDNNF-DNN, additional identity vectors (also known as iVectors) are extracted to characterize speaker characteristics.

Identity vectors (iVectors) [36] are generally used in speaker identification and represent information about the speaker and audio recording characteristics. In our context, iVectors are used as an additional input to help the network adapt to speaker characteristics and usually result in a 1% WER improvement. The iVectors used by Kaldi are similar in the sense that they are based on similar ideas as Joint Factor Analysis, where a Universal Background Model collects sufficient statistics for iVector extraction, and a Probabilis-

tic Linear Discriminant Analysis backend computes a similarity score between iVectors. Kaldi replaced the Universal Background Model based on a GMM with one considering a DNN [37].

Before the features are extracted, the data is first processed to normalize the audio and label inputs. For example, the audio is transformed to mono-channel, 8000 Hz, 16-bit PCM WAVs, and various automatic corrections are applied to the labels (for example, the punctuation marks are added, "î" inside word is replaced with "â" and vice-versa), numbers are transformed to words (for example "42" is transformed to "patruzeci şi doi"), and words are transformed to uppercase.

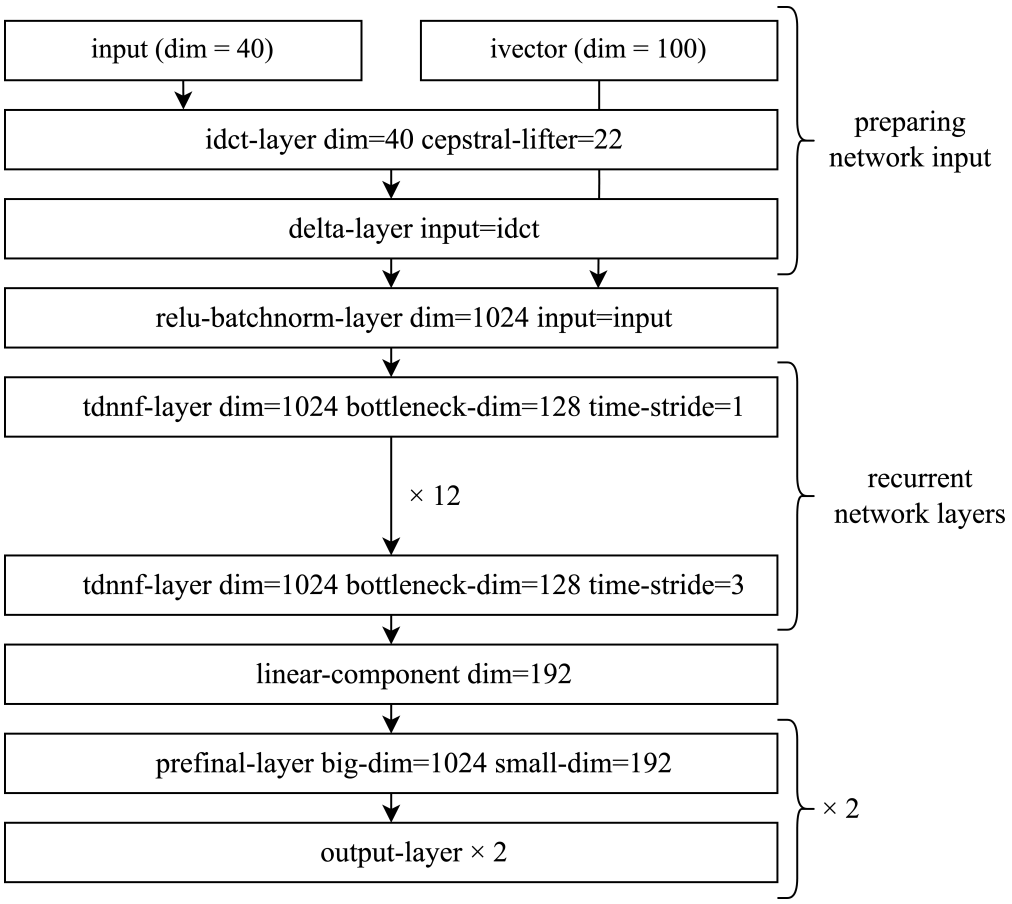


Figure 6. Model architecture for Automatic Speech Recognition system.

There are four HMM-GMM models built on top of each other, each with corresponding methods to improve accuracy and to consider larger volumes of data. Each subsequent model is used to align the training data used for the next model (see Figure 7). The first model is a simple, context-independent mono-phone model. The second model is a more complex, triphone context-dependent system trained using first (delta) and second derivatives (delta–delta or double delta) features. The third model applies Linear Discriminant Analysis (LDA) and a diagonalizing transformation known as Maximum Likelihood Linear Transform (MLLT) to the input training. The last HMM model's features are transformed one last time using a feature space maximum likelihood linear regression (fMLLR), which is known to be a Speaker Adaptive Training (SAT) method.

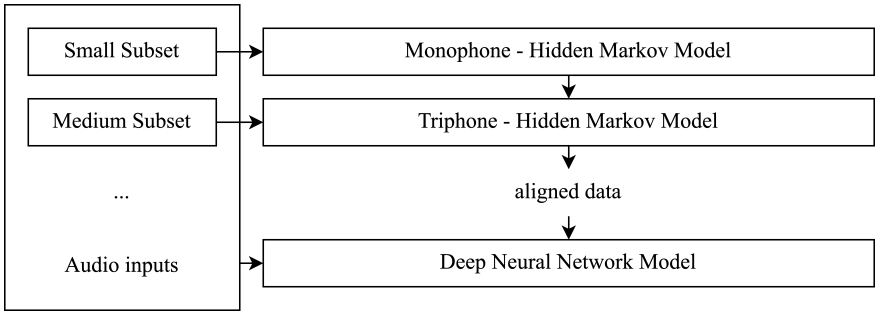


Figure 7. Training pipeline for Automatic Speech Recognition system.

The last model (i.e., our final ASR model) was trained for 12 epochs using the MFCC and iVectors of the entire training dataset. It was defined using Kaldi's XConfig domain-specific language (DSL) and employs the implementation of the neural network named "nnet3" from Kaldi (available at https://kaldi-asr.org/doc/dnn3.html, accessed on 1 December 2022; Povey et al. [38]).

3.4. Speech Emotion Recognition Model

3.4.1. Speech Pre-Processing

Our aim is to develop a system for emergency services. As a consequence, the speech data to be processed is recorded from a telephone. Hence, it passed through several coding/decoding steps in the communication network and is sampled at 8 kHz. The training and testing datasets must reflect the nature of the real speech signal. For this reason, we coded and decoded all recordings with a Full Rate GSM (GSM-FR) codec, assuming that currently, most calls to emergency services are from mobile terminals. We consider the Full Rate GSM codec as a worst-case scenario. The speech quality (measured using mean opinion score-MOS) of the decoded GSM-FR speech is lower than the quality of current codecs (e.g., Adaptive Multirate Narow Band-AMR-NB and Adaptive Multirate Wide Band-AMR-WB) [39,40]. As such, our model will also work with better codecs if satisfactory results are obtained when classifying emotional speech samples coded/decoded with the GSM Full Rate codec.

All recordings are segmented in 1-second frames, with a stride of 10 ms. For each frame, a log-spectrogram (see Figure 8) is computed. The log-spectrograms are transformed to RGB images after applying dynamic range normalization between −90 dB and −7 dB. This processing is similar to [30].

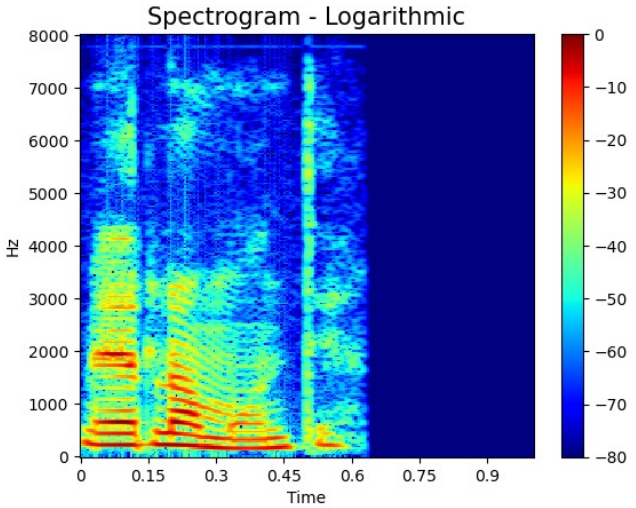


Figure 8. Sample spectrogram used as input to the neural network.

3.4.2. Neural Network Architecture

The neural network trained on top of the log-spectrogram is the well-known VGG16 [41]. VGG16 consists of a total of 21 layers, out of which 16 have learnable weights. We use the pretrained model with ImageNet and retrain the whole network (134,289,223 parameters). A block diagram is presented in Figure 9. The network's input is the log-spectrogram computed as in the previous section and reshaped to $224 \times 224 \times 3$. The output is represented by the seven classes from EmoIIT or EmoDB. The finetuning parameters are presented in Table 5.

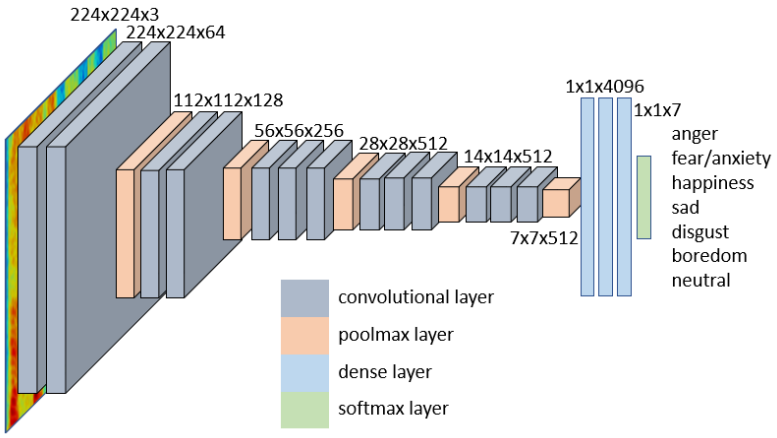


Figure 9. Neural network architecture based on VGG16.

Table 5. Fine-tuning parameters for the VGG16 experiments.

Parameter	Value
Optimization algorithm	SGDM
Mini-batch size	32
Maximum number of epochs	13
Stochastic gradient descent momentum	0.9
Initial learning rate	0.0001%
Learning rate decay	0.0001%

3.5. Deployment

The current deployment of the ODIN112 pilot consists of a PBX (Private Branch Exchange) telephone system, a microservice-based environment (including the processing services, the manager, as well as the user interface of the platform), the IVR component (as described in a previous article [33]) and an HDFS deployment used for the long-term storage of the archive.

The deployment of ODIN112 components is based on microservices. The implementation is robust using processes/threads and async I/O. We considered the interactivity and the increased potential for working with multiple simultaneous calls. The system has flexible interfaces and smooth communication as sample streams over TCP, while the communication is optimized for real-time processing.

The ODIN112 modules benefit from the entire call without having to maintain a state between requests (new samples are added to the TCP stream made for the call). The modules have the ability to return results at times independent of receiving samples.

The deployment considers only one flow per call between the manager and each processing module, and the time points (in call coordinates) are added to results directly by processing modules.

4. Results

4.1. Automatic Speech Recognition

We have trained and tested several models on various subsets of presented datasets (see Table 6. The most basic training dataset was SWARA [42]; the second training dataset was Romanian Read-Speech Corpus (RSC) [4]; the third one was the one collected by us using Echo, whereas the final one was trained using all available data. For the test experiments, we used the evaluation partition of RSC and selected 10% of the other datasets with no transcript overlap between train and test sets. For all experiments, we used the same language model to reduce its impact.

Table 6. Training and evaluation datasets duration.

Subset / Dataset	Train [hours]	Test [hours]	Total [hours]
SWARA	19	2	21
RSC	92	5	97
Echo	141	14	155
All	394	62	456

The metric used for evaluation is the Word Error Rate (WER) derived from the Levenshtein distance, which measures the difference between two sequences of words. It represents the errors in a transcript over the total number of words spoken. It can be computed as:

$$WER = \frac{S + D + I}{N} = \frac{S + D + I}{S + D + C}$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions, C is the number of correct words, and N is the number of words in the reference ($N = S + D + C$).

WER is not a perfect metric, but it is an objective way of measuring the accuracy of various speech systems. WER does not account for the type or cause of the error, whereas the transcript can still be understood by a human even if the WER is high. For example, in certain conditions, a misspelled or misrecognized word is a better output than a missing word.

4.2. Speech Emotion Recognition

We performed four experiments, as depicted in Table 7. We first tested the network on the unprocessed on both EMO-DB and EMO-IIT datasets in order to validate the network architecture. The results were similar to the state-of-the-art ones presented in other papers. The EMO-IIT results surpass previous experiments with accuracies of around 85% [24,28]. Then, we retrained the network using the GSM-like processed datasets, resulting in lower weighted accuracies by a few percent. Detailed results presented as confusion matrices are presented in Figure 10. While inspecting the confusion matrices for the EMO-IIT experiments, we notice that anger is often misclassified as sadness. We listened to some of the recordings and, indeed, some were misleading even for a human. For the EMO-DB experiments, fear was misclassified as anger and happiness but only for the GSM processed dataset.

Table 7. Results for speech emotion recognition experiments.

Corpus	Weighted Accuracy	5-Fold Cross-Validation Weighted Accuracy
EMO-DB	94.46%	92.28%
EMO-IIT	94.98%	90.24%
EMO-DB GSM	88.57%	83.86%
EMO-IIT GSM	93.46%	91.20%

(**a**) EMO-DB

(**b**) EMO-IIT

(**c**) EMO-DB GSM

(**d**) EMO-IIT GSM

Figure 10. Confusion matrices for both normal and GSM-like coded speech.

5. Discussion

5.1. Automatic Speech Recognition

Table 8 introduces the results for each model trained on a specific dataset, as well as the overarching results when the model was tested on all datasets. Each dataset obtained the best test scores when trained on the same dataset, except for the model trained on all data which performed better on the test subset of SWARA.

Table 8. Word error rates when training and testing various Romanian models. All four models were trained using only the "train" subset, whereas the evaluation was performed exclusively on the "test" subsets.

Test Set \ Train Set	SWARA [%WER]	RSC [%WER]	Echo [%WER]	All [%WER]
SWARA	3.28%	6.01%	4.42%	2.99%
RSC	13.30%	4.36%	4.40%	3.32%
Echo	37.21%	11.90%	4.97%	5.29%
All	35.38%	13.89%	4.82%	5.94%

The model trained using the Echo dataset outperforms the others to the extent that results are consistent across all test sets and the WER is between 4% and 5%; in contrast, SWARA obtains error rates between 3% and 35%, RSC between 4% and 14%, and the merged collection (i.e., All) between 3% and 6%. This argues that the model trained using the Echo dataset generalizes better, which is a direct result of the higher data quality and a larger diversity of transcripts.

The "Echo" and "All" models have similar error rates, with "All" outperforming "Echo" on SWARA and RSC, while "Echo" outperformed "All" for "Echo" and "All" datasets. Even though the difference between the two is within 1%, the model trained on all the recordings has the potential to outperform all of the other ones as the neural network used at this moment may be too shallow to learn from all the data offered as input. Nevertheless, our results are comparable to the best previously published results for the Romanian language, namely: a 3.27% WER [4] reported on the RSC dataset for a similar TDNN model, but using a Recurrent Neural Network language model; a WER of 2.79% [43] obtained also using an RNN-LM trained on the RSC and SSC – the Spontaneous Speech Corpus, summing up to over 225 h of recordings.

In addition, we rely on a simple processing of transcripts that considers only the words for training the model, without any punctuation or letter casing; this is normally handled by the language model. We will consider improving these aspects together with the language model's accuracy in future iterations of our solution.

5.2. Speech Emotion Recognition

Our results on the EMO-IIT dataset surpass the current state of the art, as shown in Table 9. To our knowledge, there are no other results for GSM-coded emotional speech recognition in Romanian. We chose to compute the weighted accuracy in order to compensate for any imbalance in the distribution of emotion classes. Our results for the EMO-DB dataset from Table 10 are similar to other current architectures. To our knowledge, the only paper with experiments on the effect of bandwidth reduction on SER is Lech et al. [30], where a pretrained AlexNet neural network is trained on log Mel spectrograms extracted from the EMO-DB corpus. Our model follows a similar approach to Lech et al. [30], though employing a more complex VGG16 type of neural network. The results from other works employing VGG16 range from an accuracy of 71% [44]) to 92% [29]. The best results on the EMO-DB corpus (95,89% accuracy) [45] were achieved using a combination of convolutional and recurrent neural networks; our result is not far behind but employs a simpler neural network model.

The train and test datasets contain high-quality entries with little noise. This is usually not the case in emergency calls. Therefore, we will consider adding noise to the

test recordings in future work. However, this is not a simple task. For this, we have to determine the minimum signal-to-noise ratio (SNR) for which a human still perceives emotions from speech and build noisy datasets with a range of SNRs around that value, followed by thorough testing of the trained models with the noisy datasets.

Table 9. Comparison of EMO-IIT results (bold marks the best model).

Works	EMO-IIT Original/GSM	Classifier	Speech Features	Test Set	Result
Feraru and Zbancioc [24]	original	DL-CNN	Mel spectrograms&MFCCs	CV-10 folds	84.48% A
Zbancioc and Feraru [28]	original	DL-CNN	Mel spectrograms	CV-10 folds	84.71% A
Our model	**original GSM**	**VGG16**	**Log spectrograms**	**CV-5 folds**	**90.24% WA 91.20% WA**

CV—crossvalidation. WA—weighted accuracy. A—accuracy.

Table 10. Comparison of EMO-DB results (bold marks the best results).

Works	EMO-IIT Original/GSM	Classifier Type	Speech Features	Test Set	Result
Popova et al. [44]	original	VGG16	Mel spectrograms	70/30 random split	71.00% A
Issa et al. [46]	original	VGG16	MFCC&chroma& Mel spectrogram& contrast&tonnetz	CV-5 folds	86.10% A
Zhao et al. [45]	**original**	**CNN-LSTM**	**Mel spectrograms**	**CV-5 folds**	**95.89% A**
Rudd et al. [29]	original	VGG16-MLP	HP & Mel spectrograms	80/10/10 random split	92.79% A
Lech et al. [30]	original GSM	AlexNet	Mel spectrograms	CV-10 folds	80.50% WA 76.80% WA
Our model	original GSM	VGG16	log Spectrograms	CV-5 folds	92.28% WA 83.86% WA

CV—crossvalidation. WA—weighted accuracy. A—accuracy. HP—harmonic percussive.

6. Conclusions

In our previous research [8], we have benchmarked various automatic speech recognition models and argued that our Kaldi is superior when it comes to its accuracy measured as WER. In this article, we have further improved our model and retrained with a far larger amount of diverse data. The accuracy levels proven by the speech technologies presented in this paper, including the automatic speech recognition one, make it a viable solution for building assistants that can help emergency services operators in making decisions.

We observed an increase in WER when the models are trained on a larger amount of data because the vocabulary is more complex and there are more variations in speech signals. However, from our lab tests that simulate real-world emergency conversations, this translates to much better transcriptions. We believe the extended dataset provides the system with a better chance of learning abstract speech rather than some very specific words or voices.

In the future, we look towards further expanding our training datasets by collecting more speech recordings through our crowdsourcing platform, Echo. Future work includes extending the neural network by increasing its parameter number in order to learn more from the training data. Finally, using richer language models and more elaborate preprocessing methods will further increase the value of the extracted features.

For emotional speech recognition, the best results in real-life applications are achieved with natural corpora. However, natural corpora are difficult to develop because recordings of actual emotional speech are not readily available and because, in some cases, the content can cause emotional harm to the listener. We are currently annotating such a corpus recorded from real-life emergency calls.

Author Contributions: Conceptualization, T.B., M.D., I.B. and F.P.; methodology, M.D. and I.B.; software, D.U., S.-A.T., I.-D.F., B.-C.M., I.A. and B.M.; validation, M.D. and I.B.; formal analysis, M.D. and F.P.; investigation, M.D., I.B. and F.P.; resources, D.U., S.-A.T., and T.B.; data curation, D.U., I.-D.F. and B.M.; writing—original draft preparation, D.U., S.-A.T., I.-D.F., B.-C.M., I.A. and B.M.; writing—review and editing, T.B., M.D., I.B. and F.P.; visualization, D.U., I.-D.F., B.-C.M. and I.A.; supervision, T.B., M.D., I.B. and F.P.; project administration, I.B.; funding acquisition, T.B., M.D., I.B. and F.P. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by a grant of the Ministry of Research, Innovation and Digitization, CNCS/CCCDI-UEFISCDI, project number PN-III-P2-2.1-SOL-2021-2-0223, within PNCDI III. This work was supported in part by OPTIM Research (POCU grant no. 62461/03.06.2022, cod SMIS: 153735).

Institutional Review Board Statement: The study was conducted in accordance with the Declaration of Helsinki and approved by the Ethics Committee of the Faculty of Automated Control and Computers, University Politehnica of Bucharest.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: The Echo dataset for Romanian is available at https://echo.readerbench.com/ (accessed on 1 December 2022).

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

ASR	Automatic Speech Recognition
CMVN	Cepstral Mean and Variance Normalization
DNN	Deep Neural Network
DSL	Domain Specific Language
fMLLR	Feature space Maximum Likelihood Linear Regression
GMM	Gaussian Mixture Model
GPU	Graphics Processing Unit
GSM	Global System for Mobile communication
HMM	Hidden Markov Model
LSTM	Long Short-Term Memory
MFCC	Mel-Frequency Cepstral Coefficients
MLLT	Maximum Likelihood Linear Transform
NLP	Natural Language Processing
RNN	Recurrent Neural Network
SAT	Speaker Adaptive Training
TDNN	Time Delay Neural Network
WER	Word Error Rate

References

1. Zicari, R.V.; Brusseau, J.; Blomberg, S.N.; Christensen, H.C.; Coffee, M.; Ganapini, M.B.; Gerke, S.; Gilbert, T.K.; Hickman, E.; Hildt, E.; et al. On Assessing Trustworthy AI in Healthcare. Machine Learning as a Supportive Tool to Recognize Cardiac Arrest in Emergency Calls. *Front. Hum. Dyn.* **2021**, *3*. [CrossRef]
2. Madsen, J.L.; Lauridsen, K.G.; Løfgren, B. In-hospital cardiac arrest call procedures and delays of the cardiac arrest team: A nationwide study. *Resusc. Plus* **2021**, *5*, 100087. [CrossRef] [PubMed]
3. Hannun, A.; Case, C.; Casper, J.; Catanzaro, B.; Diamos, G.; Elsen, E.; Prenger, R.; Satheesh, S.; Sengupta, S.; Coates, A.; et al. Deep Speech: Scaling up end-to-end speech recognition. *arXiv* **2014**, arXiv:1412.5567.
4. Georgescu, A.L.; Cucu, H.; Buzo, A.; Burileanu, C. RSC: A Romanian read speech corpus for automatic speech recognition. In Proceedings of the 12th Language Resources and Evaluation Conference, Marseille, France, 11–16 May 2020 ; pp. 6606–6612.
5. Rabiner, L.; Juang, B. An introduction to hidden Markov models. *IEEE Assp Mag.* **1986**, *3*, 4–16. [CrossRef]

6. Reynolds, D.A. Gaussian mixture models. *Encycl. Biom.* **2009**, *741*, 659–663 .
7. Lamere, P.; Kwok, P.; Gouvea, E.; Raj, B.; Singh, R.; Walker, W.; Warmuth, M.; Wolf, P. The CMU SPHINX-4 speech recognition system. In Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2003), Hong Kong, China, 6–10 April 2003; Volume 1, pp. 2–5.
8. Ungureanu, D.; Badeanu, M.; Marica, G.C.; Dascalu, M.; Tufis, D.I. Establishing a Baseline of Romanian Speech-to-Text Models. In Proceedings of the 2021 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), Bucharest, Romania, 13–15 October 2021; pp. 132–138.
9. Hochreiter, S.; Schmidhuber, J. Long short-term memory. *Neural Comput.* **1997**, *9*, 1735–1780. [CrossRef]
10. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1–11.
11. Amodei, D.; Ananthanarayanan, S.; Anubhai, R.; Bai, J.; Battenberg, E.; Case, C.; Casper, J.; Catanzaro, B.; Cheng, Q.; Chen, G.; et al. Deep speech 2: End-to-end speech recognition in English and Mandarin. In Proceedings of the International Conference on Machine Learning. PMLR, New York, NY, USA, 19–24 June 2016; pp. 173–182.
12. Radford, A.; Kim, J.W.; Xu, T.; Brockman, G.; McLeavey, C.; Sutskever, I. Robust Speech Recognition via Large-Scale Weak Supervision. *arXiv* **2022**, arXiv:2212.04356.
13. Baevski, A.; Zhou, Y.; Mohamed, A.; Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 12449–12460.
14. Povey, D.; Ghoshal, A.; Boulianne, G.; Burget, L.; Glembek, O.; Goel, N.; Hannemann, M.; Motlicek, P.; Qian, Y.; Schwarz, P.; et al. The Kaldi speech recognition toolkit. In Proceedings of the IEEE 2011 workshop on Automatic Speech Recognition and Understanding, Waikoloa, HI, USA, 11–15 December 2011.
15. Heafield, K. KenLM: Faster and smaller language model queries. In Proceedings of the Sixth Workshop on Statistical Machine Translation, Edinburgh, UK, 30–31 July 2011; pp. 187–197.
16. Eckman, P.; Friesen, V., W.; Ellsworth, P. *Emotion in the Human Face Guidelines for Research and an Integration of Findings Volume 11 in Pergamon General Psychology Series*; Elsevier Inc.: Amsterdam, The Netherlands, 1972.
17. Busso, C.; Bulut, M.; Lee, C.C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J.N.; Lee, S.; Narayanan, S.S. IEMOCAP: Interactive emotional dyadic motion capture database. *Lang. Resour. Eval.* **2008**, *42*, 335–359. [CrossRef]
18. Burkhardt, F.; Paeschke, A.; Rolfes, M.; Sendlmeier, W.; Weiss, B. A database of German emotional speech. In Proceedings of the 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, 4–8 September 2005; Volume 5, pp. 1517–1520. [CrossRef]
19. Pichora-Fuller, M., K.; Dupuis, K. Toronto emotional speech set (TESS). In *Scholars Portal Dataverse*; University of Toronto, Toronto, ON, Canada, 2020. [CrossRef]
20. Engberg, S., I.; Hansen, A.V.; Andersen, O.; Dalsgaard, P. Design, recording and verification of a danish emotional speech database. In Proceedings of the Eurospeech, Rhodes, Greece, 22–25 September 1997.
21. Costantini, G.; Iaderola, I.; Paoloni, A.; Todisco, M. EMOVO Corpus: an Italian Emotional Speech Database. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, 26–31 May 2014; European Language Resources Association (ELRA): Reykjavik, Iceland, 2014; pp. 3501–3504.
22. Martin, O.; Kotsia, I.; Macq, B.; Pitas, I. The eNTERFACE'05 Audio-Visual Emotion Database. In Proceedings of the 22nd International Conference on Data Engineering Workshops (ICDEW'06), Washington, DC, USA, 3–7 April 2006.
23. Kossaifi, J.; Tzimiropoulos, G.; Todorovic, S.; Pantic, M. AFEW-VA database for valence and arousal estimation in-the-wild. *Image Vis. Comput.* **2017**, *65*, 23–36. [CrossRef]
24. Feraru, M.; Zbancioc, M.D. Emotion Recognition Results using Deep Learning Neural Networks for the Romanian and German Language. In Proceedings of the 2020 International Conference on e-Health and Bioengineering (EHB), Iasi, Romania, 29–30 October 2020; pp. 1–4. [CrossRef]
25. Wani, T.M.; Gunawan, T.S.; Qadri, S.A.A.; Kartiwi, M.; Ambikairajah, E. A Comprehensive Review of Speech Emotion Recognition Systems. *IEEE Access* **2021**, *9*, 47795–47814. [CrossRef]
26. Badshah, A.M.; Ahmad, J.; Rahim, N.; Baik, S.W. Speech emotion recognition from spectrograms with deep convolutional neural network. In Proceedings of the 2017 International Conference on Platform Technology and Service (PlatCon), Busan, Republic of Korea, 13–15 February 2017.
27. Mirsamadi, S.; Barsoum, E.; Zhang, C. Automatic speech emotion recognition using recurrent neural networks with local attention. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017.
28. Zbancioc, M.D.; Feraru, S.M. Emotion Recognition for Romanian Language Using MFSC Images with Deep-Learning Neural Networks. In Proceedings of the 2021 International Conference on e-Health and Bioengineering (EHB), Iasi, Romania, 18–19 November 2021; pp. 1–4. [CrossRef]
29. Rudd, D.H.; Huo, H.; Xu, G. Leveraged Mel Spectrograms Using Harmonic and Percussive Components in Speech Emotion Recognition. In *Advances in Knowledge Discovery and Data Mining*; Gama, J.; Li, T.; Yu, Y.; Chen, E.; Zheng, Y.; Teng, F., Eds.; Springer International Publishing: Cham, Switzerland, 2022; pp. 392–404.
30. Lech, M.; Stolar, M.; Best, C.; Bolia, R. Real-Time Speech Emotion Recognition Using a Pre-trained Image Classification Network: Effects of Bandwidth Reduction and Companding. *Front. Comput. Sci.* **2020**, *2*, 14. [CrossRef]

31. Sun, L.; Fu, S.; Wang, F. Decision tree SVM model with Fisher feature selection for speech emotion recognition. *EURASIP J. Audio Speech Music. Process.* **2019**, *2019*, 2. [CrossRef]
32. Park, D.S.; Chan, W.; Zhang, Y.; Chiu, C.C.; Zoph, B.; Cubuk, E.D.; Le, Q.V. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. *arXiv* **2019**, arXiv:1904.08779.
33. Mocanu, B.C.; Filip, I.D.; Ungureanu, R.D.; Negru, C.; Dascalu, M.; Toma, S.A.; Balan, T.C.; Bica, I.; Pop, F. ODIN IVR-Interactive Solution for Emergency Calls Handling. *Appl. Sci.* **2022**, *12*, 10844. [CrossRef]
34. Masala, M.; Ruseti, S.; Dascalu, M. Robert–a romanian bert model. In Proceedings of the 28th International Conference on Computational Linguistics, Barcelona, Spain, 8–13 December 2020; pp. 6626–6637.
35. Mermelstein, P. Distance measures for speech recognition, psychological and instrumental. *Pattern Recognit. Artif. Intell.* **1976**, *116*, 374–388.
36. Dehak, N.; Kenny, P.J.; Dehak, R.; Dumouchel, P.; Ouellet, P. Front-end factor analysis for speaker verification. *IEEE Trans. Audio Speech Lang. Process.* **2010**, *19*, 788–798. [CrossRef]
37. Snyder, D.; Garcia-Romero, D.; Povey, D. Time delay deep neural network-based universal background models for speaker recognition. In Proceedings of the 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU), Scottsdale, AZ, USA, 13–17 December 2015; pp. 92–97.
38. Povey, D.; Zhang, X.; Khudanpur, S. Parallel training of DNNs with natural gradient and parameter averaging. *arXiv* **2014**, arXiv:1410.7455.
39. Vary, P.; Hellwig, K.; Hofmann, R.; Sluyter, R.; Galand, C.; Rosso, M. Speech codec for the European mobile radio system. In Proceedings of the ICASSP-88, International Conference on Acoustics, Speech, and Signal Processing, New York, NY, USA, 11–14 April 1988; Volume 1, pp. 227–230. [CrossRef]
40. Holma, H.; Melero, J.; Vainio, J.; Halonen, T.; Makinen, J. Performance of adaptive multirate (AMR) voice in GSM and WCDMA. In Proceedings of the The 57th IEEE Semiannual Vehicular Technology Conference, 2003. VTC 2003-Spring., Jeju, Republic of Korea, 22–25 April 2003; Volume 4, pp. 2177–2181. [CrossRef]
41. Simonyan, K.; Zisserman, A. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv* **2014**, arXiv:1409.1556.
42. Stan, A.; Dinescu, F.; Ţiple, C.; Meza, Ş.; Orza, B.; Chirilă, M.; Giurgiu, M. The SWARA speech corpus: A large parallel Romanian read speech dataset. In Proceedings of the 2017 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), Bucharest, Romania, 6–9 July 2017; pp. 1–6.
43. Georgescu, A.L.; Cucu, H.; Burileanu, C. Kaldi-based DNN Architectures for Speech Recognition in Romanian. In Proceedings of the 2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), Timisoara, Romania, 10–12 October 2019; pp. 1–6. [CrossRef]
44. Popova, A.S.; Rassadin, A.G.; Ponomarenko, A. *Emotion Recognition in Sound*; Springer: Berlin/Heidelberg, Germany, 2017.
45. Zhao, J.; Mao, X.; Chen, L. Speech emotion recognition using deep 1D & 2D CNN LSTM networks. *Biomed. Signal Process. Control.* **2019**, *47*, 312–323.
46. Issa, D.; Demirci, M.F.; Yazıcı, A. Speech emotion recognition with deep convolutional neural networks. *Biomed. Signal Process. Control* **2020**, *59*, 101894. [CrossRef]