SPEECH-BASED ACTIVITY RECOGNITION FOR MEDICAL TEAMWORK

by

JALAL NAZAR ABDULBAQI

A dissertation submitted to the

School of Graduate Studies

Rutgers, The State University of New Jersey

In partial fulfillment of the requirements

For the degree of

Doctor of Philosophy

Graduate Program in Electrical and Computer Engineering

Written under the direction of

Ivan Marsic

And approved by

____

____

____

____

New Brunswick, New Jersey

October 2020

ABSTRACT OF THE DISSERTATION

Speech-Based Activity Recognition for Medical Teamwork

by Jalal Nazar Abdulbaqi
Dissertation Director: Ivan Marsic

Activity recognition is the process by which one or more people's actions and their environment are observed and analyzed to infer their activities. The activity recognition task includes recognizing the activity type and estimates its progression stage, from planning, through performance, to evaluation. This task is usually achieved by using different types of sensor modalities such as video, radio frequency identification (RFID), and medical device signals. To our knowledge, the speech data representing the verbal communication between individuals has not been used for activity recognition. In medical teamwork, some activities are conducted through verbal communication between the medical team. Consequently, for these activities, speech data can provide better information than video or other sensors. On the other hand, speech data present challenges that include fast and concurrent talking (known as “cocktail party problem”), as well as the ambient noise. Therefore, achieving speech-based activity recognition requires dealing with these challenges as well as finding the optimal model architecture for activity recognition. In this research, we develop deep neural networks that use and enhance the utterance-level speech stream to predict medical activity type.

Our speech-based approach to recognize team activities is developed in the context of trauma resuscitation. We first analyzed the audio recordings of trauma resuscitations in terms of activity frequency, noise-level, and activity-related keyword frequency to determine the dataset characteristics. We next evaluated different audio-preprocessing parameters (spectral feature types and audio channels) to find the optimal configuration. We then introduced a novel neural network to recognize the trauma activities using a modified VGG network that extracts features from the audio input. The output of the modified VGG network is combined with the output of a network that takes keyword text as input, and the combination is used to generate activity labels. We compared our system with several baselines and performed a detailed analysis of the performance results for specific activities. Our results show that our proposed architecture that uses Mel-spectrum spectral coefficients features with a stereo channel and activity-specific frequent keywords achieve the highest accuracy and average F1-score.

We further propose an extensive analysis of keyword labeling. We investigated two approaches to create the keyword list: by the number of their existence in the dataset and by computing the keyword sensitivity for each activity. Besides, we examine using a different number of keywords per activity to find the optimum number. Also, we categorize the keywords based on their relationship to the medical activities to find a nonvaluable keyword to remove. This analysis assists us to increase the performance significantly.

We present a broad analysis of the multimodal network for trauma activity recognition. That includes the audio, keyword, and fusion modules. We design and evaluate different networks and learning approaches for audio, keyword, and fusion modules. For the audio network, we design and examine two networks: the first network uses a convolutional neural network (CNN), while the second network, we replace several frontend CNN layers with two recurrent neural networks (RNN) to track the temporal information in the sequential speech recordings. For the keyword network, we propose five networks for evaluation to extract the features from the transcript keyword input. For the fusion, we applied and examine two fusions approaches early fusion and late fusion. Evaluation results show substantial improvement in the accuracy and the f1-score over the baseline

An important challenge that affects the performance introduced speech-based activity recognition approach, is the speech quality in general and the non-stationary noise in particular. An efficient speech enhancement system is required to address this issue. Most current speech enhancement models use spectrogram features that require an expensive transformation and result in phase information loss. Previous work has overcome these issues by using convolutional networks to learn the temporal correlations across high-resolution waveforms. These models, however, are limited by memory-intensive dilated convolution and aliasing artifacts from upsampling. We introduce an end-to-end, fully recurrent neural network for single-channel speech enhancement. Our network is structured as an hourglass-shape that can efficiently capture long-range temporal dependencies by reducing the feature resolution without information loss. Also, we used residual connections to prevent gradient decay over layers and improve the model generalization. Experimental results show that our model outperforms state-of-the-art approaches in six quantitative evaluation metrics.

DEDICATION

To my wonderful parents

Laeqa Fadhil Shawket and ***Nazar Abdulbaqi Mahmood***

To my lovely wife

Sumayah Abid Al-Mahmood

To my amazing kids

Ibrahim**, **Mohammed**,* and ***Abdullah

To my supportive brother and sister

Hazim Nazar Abdulbaqi and ***Sundus Nazar Abdulbaqi***

ACKNOWLEDGMENTS

I would like to express my grateful and deepest appreciation to my Advisor Professor **Ivan Marsic** for his inspiration, mentoring and support through my research. Also, I would like to thank my dissertation defense committee members: Professors **Zoran Gajic**, Professor **Predrag Spasojevic**, and Professor **Aleksandra Sarcevic** for their generous time and valuable comments.

I would like to express my gratitude to my colleagues in Signal Processing Lab: Dr. **Yue Gu**, **Zhichao Xu**, Dr. **Xinyu Li**, **Chenyang Gao** and **Yanyi Zhang** for their generous assistance and effective discussions during my work in the lab.

I would like to recognize the stuff in the Electrical and Computer Engineering at Rutgers University, particularly, the wonderful graduate director Professor **Zoran Gajic** for his support over all my study years at the department. Likewise, I would like to thank our marvelous department stuff: **Christy Lafferty**, **Arletta Hoscilowicz**, **John Scafidi**, **Christopher Reid** and **Pamela Heinold** for their help to complete my work.

I would Like to thank the stuff in the School of Graduate Studies, remarkably a deep grateful to Mrs. **Alex Bachmann** for her endless support and assistant.

I would like to thank to the stuff in the Children's National Medical Center, Washington, DC, especially Professor **Randall Burd**, MD for his generous and thoughtful mentoring. Also, special thanks to **Megan Cheng**, **Yinan Zheng**, **Adesh Ranganna** and **Emily Alberto** for their efforts to prepare the dataset for this research.

I would like to acknowledge our research partners at Drexel University, Philadelphia, PA, especially Professor **Aleksandra Sarcevic** for her supporting efforts. Also, I would like to thank our remote colleagues: **Swathi Jagannath**, **Leah Kulp**, and **Angela Mastrianni** for their valuable assistance.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Overview

Activity recognition is the process by which one or more people's actions and their environment are observed and analyzed to infer their activities. Activity recognition plays an important role in medical setups such as trauma resuscitation to reduce medical errors. Predicting the activities from different visual, audio, and sensory data resources assist the medical team to minimize the miscommunication and medical mistakes. However, activity recognition of a dynamic medical scheme is challenging owing to fast and concurrent events as well as noisy environments. Current research relies on video radio frequency identification (RFID), and signals from devices to identify the medical activity and its phase [1]–[3]. Visual and RFID resources provide free-hand, cheap, and high accuracy predictions for certain activities. However, these data resources have their limitations. Visual data resource occasionally relies on RBG camera, which can violate patient privacy. Also, several activities cannot be predicted because either there are no visual actions that can be observed or the activity is not related to a specific medical device that can be tagged with RFID. The verbal communication among the medical team is a valuable source of data for activity recognition for different reasons: first, certain activities are verbal-reliant, which can be predicated only by observing the speech data. Second, previous studies showed that fusing the speech with the video, RFID or transcripts increases activity recognition accuracy [5], [6]. Finally, a study [4] found that medical experts can predict resuscitation activities with 87% accuracy using only the verbal communication transcripts. However, speech-based activity recognition facing certain challenges such as rapid and simultaneous speaking along with the noisy environment.

Hypothetically, predicting the activities from the speech dataset can be implemented using two approaches: direct recognition and using automatic speech recognition (ASR). Direct recognition is implemented by feeding acoustic features directly into the model to predict the activities. We employ this approach using our baseline model described in chapter two. Our evaluation results indicate poor performance with 30.8% accuracy and 0.231 in average F1-Score. The second approach is implemented in two stages by recognizing the speech utterances transcripts using the ASR, then feeding this transcript into another model to predict the activities. Although, the study [4] shows that activity recognition from the transcripts achieves 69.1 accuracy and 0.67 average F1-Score, our ASR evaluation outcomes on the resuscitation dataset using two different architectures: attention-based seq2seq [5] and N-gram [6] achieved a high word error rate (WER) with 100.3 and 75.8 WER, respectively. Therefore, we propose a new approach by introducing the keywords as an additional input.

In this work, we introduce the keyword approach as a replacement for the previously described methods. Our method relies on using one chosen word from the utterance as an additional input to the acoustic feature to predict the activities. We think that recognizing one word from the transcript is much simple than predicting the whole utterance. Developing this approach passes through three stages. Dataset preprocessing and features representation format, baseline model design and evaluation, and improving the architecture modules and keyword labeling.

In the first, we analyze the trauma resuscitations dataset in terms of activity frequency, noise-level, and activity-related keyword frequency. We count 30 high-level activities, which are not distributed uniformly in the dataset. For example, there are 836 utterances labeled with extremity (E), while there are only 13 utterances labeled with relieve obstruction (RO). Therefore, we choose the first five high frequent activities and we label all other activities with the "OTHERS" label. Then, we evaluate the noise-level subjectively by listening to each trauma case individually to determine the clarity of medical team speech especially when they report the patient status, which indicates the activities. Finally, we analyze the keywords list that can assist the prediction of each activity. In this stage, we present the most-frequent keywords per activity approach to creating the keyword list, but later in the third stage, we create the keyword list by calculating the sensitivity for each keyword per activity. Next, we evaluated two audio features preprocessing parameters: spectral feature types and audio channels. We present an evaluation comparison between two acoustic feature spectrogram representation: Mel-frequency cepstral coefficients (MFCC) and Mel-frequency spectral coefficients (MFSC). Our results show that MFSC performance is leading over MFSC with and without using the temporal derivatives (deltas). Our audio dataset is recorded using two microphones. This provides us the possibility to examine five different input channels configurations. Assessment findings indicate that combining the two channels always better than using one channel alone. Also, feeding the two channels into the model deliver better performance than getting the average of them.

We then introduced a novel neural network to recognize the trauma activities using a modified VGG network that extracts features from the audio input. The output of the modified VGG network is combined with the output of a network that takes keyword text as input, and the combination is used to predict activities. We provide an evaluation comparison between our audio module and several the-state-of-art classification models in which our model outperforms other models.

In the next stage, we propose a wide analysis of keyword labeling. We investigated two approaches to create the keyword list: the current approach by the number of their existence in the dataset and the new approach by computing the keyword sensitivity for each activity. The new sensitivity approach outperforms the current keyword frequency in the evaluation results. Also, we examine using a different number of keywords per activity to find the optimum number. The evaluation results show that between 30 to 40 keywords per activity have higher accuracy concerning less than 30 or more than 40 keywords per activity. Also, filter the created keyword list based on their relationship to the medical activities by removing the nonrelated keywords., which assists us to increase the performance significantly.

In addition to the keyword analysis, we present a broad analysis for the activity recognition multimodal network modules for the trauma resuscitation, which include audio, keyword, and fusion modules. We design and evaluate different networks and learning approaches for audio, keyword, and fusion modules. For the audio network, we design and examine two networks: the first network uses a convolutional neural network (CNN), while the second network, we replace several frontend CNN layers with two recurrent neural networks (RNN) to track the temporal information in the sequential speech recordings. The new RNN-CNN network surpasses the previous CNN network in the evaluation outcomes. For the keyword network, we propose five different networks for evaluation to extract the features from the transcript keyword input. Two models used fully connected networks (FCN); two models used CNN, and the last one used RNN. The difference between the two versions of the FCN and the two of CNN is in the network depth. The evaluation results showed that, in general, fully connected network networks outperformed both RNN and CNN networks and the deeper FCN achieved the highest accuracy and the F1-Score. For the fusion, we applied and examine two fusions approaches early fusion and late fusion. Late fusion good improvement over the early fusion in the evaluation results in terms of accuracy and F1-Score.

Ambient noise is one of the main factors that degrade the speech-based activity recognition performance. Therefore, we introduce a novel speech enhancement system to determine this challenge. Most current speech enhancement models use spectrogram features that require an expensive transformation and result in phase information loss. Previous work has overcome these issues by using convolutional networks to learn the temporal correlations across high-resolution waveforms. These models, however, are limited by memory-intensive dilated convolution and aliasing artifacts from upsampling. We introduce an end-to-end, fully recurrent neural network for single-channel speech enhancement. Our network is structured as an hourglass-shape that can efficiently capture long-range temporal dependencies by reducing the feature resolution without information loss. Also, we used residual connections to prevent gradient decay over layers and improve the model generalization. Experimental results show that our model outperforms state-of-the-art approaches in six quantitative evaluation metrics.

1.2 Organization

These consist of three chapters. In chapter two, we introduce an extensive analysis of the dataset and feature preprocessing. We examine and evaluate different input preprocessing setup. Also, we present the first design for speech-based activity recognition, which we introduce the new keyword approach. In chapter three, we provide deeper analysis for keyword labeling algorithm and multimodal architecture. In chapter four, we present the speech enhancement model for non-stationary noises. We provide a literature comparison between the previous research. Finally, a conclusion chapter that summarizes the thesis.

1.3 Contribution

Our contribution in this research can be concluded as follows:

1 An analysis of real medical teamwork (trauma resuscitation) dataset characteristics to determine the constraints related to speech-based activity recognition.

2 Audio preprocessing analysis to find the optimal parameters for designing the network.

3 Design of an audio classification network and comparison of its performance to the state-of-the-art classification models using a trauma resuscitation audio dataset.

4 Present a new keyword-based network approach for activity recognition that combines the audio stream and the most frequent words from the input transcript.

5 An end-to-end multimodal architecture that uses speech utterances and related keywords to recognize trauma resuscitation activities.

6 A keywords analysis to estimate each keyword weight for activity prediction.

7 An analysis for each multimodal module: audio, keyword, and fusion networks.

8 Introducing a new speech enhancement for non-stationary noises to improve the quality of the medical audio recordings.

CHAPTER 2

DATA ANALYSIS AND MODEL ARCHITECTURE

2.1 Overview

Activity recognition of a dynamic medical process such as trauma resuscitation is challenging because of fast and concurrent work as well as a noisy environment. Several current research approaches rely on video, RFID, and signals from medical devices to identify the medical activity type and its stages [1]–[3]. However, to our knowledge, there have been no approaches that rely on the speech from verbal communication of the team. Video and RFID data cannot provide information to recognize certain activities. For instance, in the trauma resuscitation, Glasgow coma score calculation (GCS) and airway assessment (AA) activities rely on visual examination or talking to the patient and can be recognized only based on verbal communication. We asked three medical experts to rate different modalities (speech, video, and RFID-tagged objects) as the best source for recognizing different ongoing resuscitation activities. In Table 2.1, we averaged their ratings for four activities for which speech was rated the highest as a prediction source had they been asked to do activity recognition. Besides, a study [7] found that medical experts can predict resuscitation activities with 87% accuracy using only the verbal communication transcripts. Furthermore, previous studies showed that fusing the speech with the video, RFID or transcripts increases activity recognition accuracy [4], [8].

Activity	Audio (%)	Video (%)	RFID tag
GCS Calculation	80	7.5	Non
Airway Assessment	80	20	Non
Medications	80	20	Partial
CPR	65	45	Partial

Table 2.1: Activities for which the medical expert rated highest speech as the modality for activity recognition.

We present a speech-based activity recognition design for dynamic medical teamwork and empirical evaluation. Our approach is based on using one representative keyword from the input utterance to the activity recognition network, in addition to the audio stream. This keyword belongs to the most frequent words list that has been calculated for each activity type. Besides, to determine the challenges related to system design and dataset limitations, we determined the dataset characteristics related to the activities (e.g. activity frequency, noise-level, and word frequency for each activity). Then, we analyzed different audio preprocessing parameters, such as feature types and input channels to find the best input feature setup. Using these findings, we designed an audio classification network based on the VGG model [9]. We evaluated our audio network and compared its performance with several state-of-the-art classification networks using the trauma resuscitation dataset. Finally, we evaluated our keyword-based network design using different settings for the network layers. We found that a keyword-based approach to activity recognition performed better than relying on manually-generated transcripts. The results show that our new keyword-based design increased the accuracy and the average F1-score by 3.6% and 0.184 respectively compared to our audio network alone.

2.2 Related Work

In recent years, activity recognition for clinical purposes has been grown quickly. Most of the current research relies on the sensors and visual modalities such as passive RFID and the videos, respectively. While there are a few kinds of research works on audio and verbal information.

RFID-based activity recognition considered an object-use detection problem. Early work compares different machine learning approaches as a binary classifier to predict the medical object motion that related to certain activities [10]. A different strategy to place the RFID tags shows an improvement in the activity recognition accuracy [11], [12]. Recently, employing a convolutional neural network (CNN) as a multi-class classifier outperform the previous approaches [2]. Although, RFID has many advantages such as its small, cheap, and battery-free, the devices' radio noise and the limited number of activities that can be predicted to limit its accuracy and scalability.

Visual-based activity recognition exploits the visual data from RGB or depth camera to map the medical team movement and actions into activities. Early research uses a single camera video recording with the Markov Logic Network model to predict the activities [1]. Recently, deep learning has been applied to visual-based activity recognition. The convolutional neural network has been applied for video classification using time-stacked frames with a slow fusion network to process the short-range temporal association of activities [13]. To address the short-range temporal limitation, a long short term memory network (LSTM) has been integrated with the VGG network with a region-based technique to generate an activity mask [14]. Despite the decent progress in utilizing virtual data, it has several limitations. The RBG camera can violate patient privacy and as in the RFID, not all activities can be predicted by tracking the medical team movement and actions.

Text-based activity recognition employs the transcript of the verbal communication between the medical team to predict the activity type. Recent research applies a multi-head attention architecture [15] to predict a speech-reliant activity from the transcripts and the environmental sound [4]. The drawback of this approach is that the text modality is not available automatically, which required additional automatic speech recognition (ASR) to provide the transcripts.

The audio modality was used as an auxiliary to other modalities in works [4], [8]. These papers analyzed the audio ability to improve the accuracy of the activity recognition. In [8], the authors built a multimodal system to recognize concurrent activities by using multiple data modalities: depth camera video, RFID sensors, and audio recordings. Each modality processed and the features extracted by a separate convolutional neural network (CNN), and then all of them fused using Long Short-Term Memory (LSTM) network to the final decision layer. They did not provide quantitative analysis to distinguish the difference between each modality performance. In [4], the authors created a multimodal transformer network to process the transcribed spoken language and the environmental sound to predict the trauma activities. The quantitative analysis showed an average accuracy of 36.4 when using only audio, and the accuracy increased to 71.8 when using both modalities.

2.3 Dataset Collection and Characteristics

The dataset was collected during 86 trauma resuscitations in the emergency room at a pediatric teaching hospital in the U.S. Mid-Atlantic region between December 2016 – May 2017. We obtained approvals from the hospital’s Institutional Review Board (IRB) before the study. All data generated during the study were kept confidential and secure following IRB policies and Health Insurance Portability and Accountability Act (HIPAA). The audio data was recorded using two fixed NTG2 Phantom Powered Condenser shotgun microphones. These microphones pointed in two locations where the key members of a trauma team normally stand. The recordings were manually transcribed and each sentence was assigned the activity label by trauma experts. In this section, we present an analysis of the following three characteristics of the dataset that can affect the activity recognition outcome: activity frequency, noise level, and word frequency for each activity.

The fine-grained activities have been grouped into 30 high-level categories. Different categories occurred with different frequencies, which is the total number of utterances that include a given activity category for the 86 resuscitations cases (Table 2.2). As seen, the activities are not distributed uniformly over the dataset utterances. Some activities occurred

#	**Activity**	**Code**	**Utterances**
1	Extremity	E	836
2	Back	BK	701
3	GCS Calculation	GCS	610
4	Face	F	514
5	Circulation Control	CC	407
6	Log Roll	LOG	389
7	C-Spine	CS	380
8	Medications	MEDS	358
9	Pulse Check	PC	289
10	Blood Pressure	BP	256
11	Ear Assessment	EAR	246
12	Eye Assessment	EY	246
13	Exposure Control	EC	220
14	Abdomen Assessment	A	208
15	Breathing Assessment	BA	206
16	Airway Assessment	AA	197
17	Head	H	175
18	Exposure Assessment	EA	174
19	CPR	CPR	160
20	Chest Palpation	CP	150
21	Breathing Control	BC	137
22	Pelvis Assessment	PE	122
23	LEADS	LEADS	116
24	Endotracheal Tube Endorsement	ET	109
25	Neck Assessment	NECK	96
26	Intubation	I	50
27	Genital Assessment	G	44
28	NGT	NGT	30
29	Bolus	B	18
30	Relieve Obstruction	RO	13
	Total activity-labeled utterances		7457

Table 2.2: Resuscitation activities with most utterances.

very frequently, while others were rare. There are several reasons for this variation. First, the length of conversation between the medical team is different for each activity. Some activities require several inquiries and reports, while other activities may have s single sentence to report the patient's status. Second, each patient required different evaluation and management activities based on the patient's injury, demographics, and medical situation.

Finally, as mentioned above, the activity categories are a high-level group that sometimes includes several low-level activities (Table 2.3). Hence, when an activity group (e.g. Extremity Assessment) has several low-level activities, this tends to correspond to more verbal communication among the medical team. As a result of this non-uniform activity distribution, it is hard to train a neural network model for the activities that had associated least-frequent utterances even for activities that cannot be recognized from other modalities (e.g. Airway Assessment) because of insufficient data to train and evaluate the model. Therefore, we choose the top five activities that had associated most-frequent

High-Level Activity	**Low-level Activity**
GCS Calculation	Verbalized Motor Assess Verbal Assess Eye Assess
Extremity Assessment	Right Upper Left Upper Right Lower Left Lower
Medications	Medications
Airway Assessment	Airway Assessment
CPR	Chest comp Shock Defib pads ID

Table 2.3: Four high-level activities and their related low-level activities.

utterances from Table 2.2 for our experiments: Extremity, Back, GCS Calculation, Face, and Circulation Control. All other utterances that do not belong to these activity categories are assigned to the "OTHER" category.

The second important dataset parameter that can influence the recognition network performance is the ambient noise. The resuscitation environment presents several challenges to speech-based activity recognition. Concurrent speakers ("cocktail party" problem), rapid speech and ambient noise adversely affect the speech quality and reduce activity recognition accuracy. To estimate the clarity of the medical team speech, we performed a subjective evaluation of the trauma resuscitation dataset. In this evaluation, we categorized the noisiness of audio recordings into three levels based on the human ability to understand the reports of patient vital signs and examination results. Three medical experts worked on this assessment listening to the 86 resuscitation cases. Each case had been labeled with one of the three noise categories (low, medium, and high) and the average is shown in Table 2.4. As seen, about 65% of cases were labeled as low-noise, while about 19% and 16% were labeled as a medium- and high-level, respectively. Thus, about 35% in our dataset are either unintelligible or it is hard to understand what the medical team said during the resuscitation, which is challenging for the neural network performance. To study the effect of the ambient noise on the recognition of our selected activities, we calculated the number of noisy cases for each chosen activity (Figure 2.1).

Noise Level	**Number of Cases**
High	14
Medium	16
Low	56
Total	86

Table 2.4: A subjective evaluation of noise for all 86 resuscitation cases by three raters.

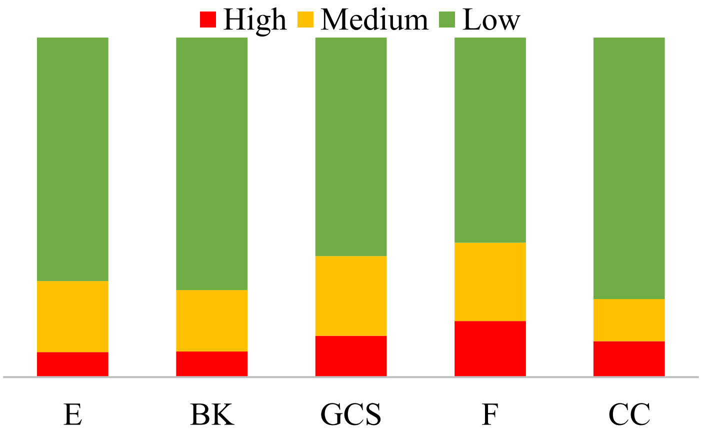

Figure 2.1: Cases noise-level distribution for each activity.

Figure 2.1 shows the fraction of the resuscitation cases by their noise level for each activity. As seen, the noise is distributed almost uniformly among the activities in our experiment. Therefore, it is not expected to affect some activities more than others in terms of prediction accuracy.

The patient medical status keywords are the most important information of the medical team verbal communication in the trauma resuscitation, which sometimes indicates the activity explicitly (e.g. GCS in Figure 2.2). To find the priority of the keywords concerning the related activities, first, we filtered most of the stop words then, we calculate the most-frequent words for each activity (Figure 2.2). As seen, most of the shown keywords either have a direct relationship with the activity (e.g. "spine" for BACK) or have indirect meaning such as the body position (e.g. right or left for extremity). Also, we can see that

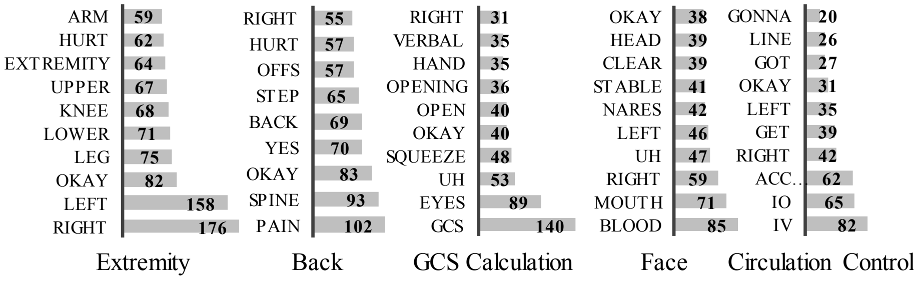

Figure 2.2: The most frequent unique words for each activity.

several words have no relationship with the activities, but they are repeated many times as a part of the inquiry response (e.g. “okay”) or just part of a repeated sentence (e.g. “get”). However, our intuition is that as long as these words are repeated frequently for certain activity then these words are valuable for the neural network to predict the correct activity. Hence, we hypothesize that these keywords can be combined with the audio stream and fed into the proposed network to increase the activity recognition accuracy. Extracting these keywords can be done automatically using a word-spotting model. We believe that extracting such keywords is easier and more efficient than recognize the whole utterance transcripts using an ASR system. In this work, we evaluated the concept of combining keywords and the audio stream to improve activity recognition performance. Although evaluating a word-spotting model is not part of this work, we will consider that in our future research.

2.4 Data Preprocessing and Configuration

The main input data is the utterance-level audio stream. Also, we consider using one keyword from the most frequent words list as an additional input. The keyword input is encoded as a one-hot one-dimension vector. Whereas, the audio stream is converted into a spectrogram. Spectrogram representation reduces the dimension of the data and provides better information representation [16]. In this section, we present a description for the data preprocessing and an investigation for two parameters variation effect on the activity recognition outcomes: feature type and input channels.

The keyword feature represented as a one-hot vector of size 78 to represent the total 60 words list (10 keywords per activity). The one-hot vector size had been incremented by 0.3 to reduce the one-hot hash function collision probability. The audio recordings were sampled at 16MHz. We used 40 filter banks for the short-time Fourier transform with a 2048 window, 25% overlap, and Hann window type. The audio stream utterances had different time lengths (Figure 2.3). The average utterance time duration was 2.42 seconds with a standard deviation of 2.28. Our neural network required a fixed input length, which can be implemented in several ways. First, we could choose a small input size that most of the utterances have such as 1-2 sec or 2-3 sec, but this would reduce the total number of samples. Second, we could specify a fixed length such as the average value and then truncate all the longer utterances, but our experiments showed that the lost information would significantly reduce the performance. Therefore, we resized all the utterance lengths to be 20 seconds by zero-padding the beginning and end of each utterance. The final feature map length was 600. Following the work [4], we segmented the input feature map into 10 frame sub-maps to avoid processing distant audio frames. The input sample shape for every single channel was (60, 40, 10).

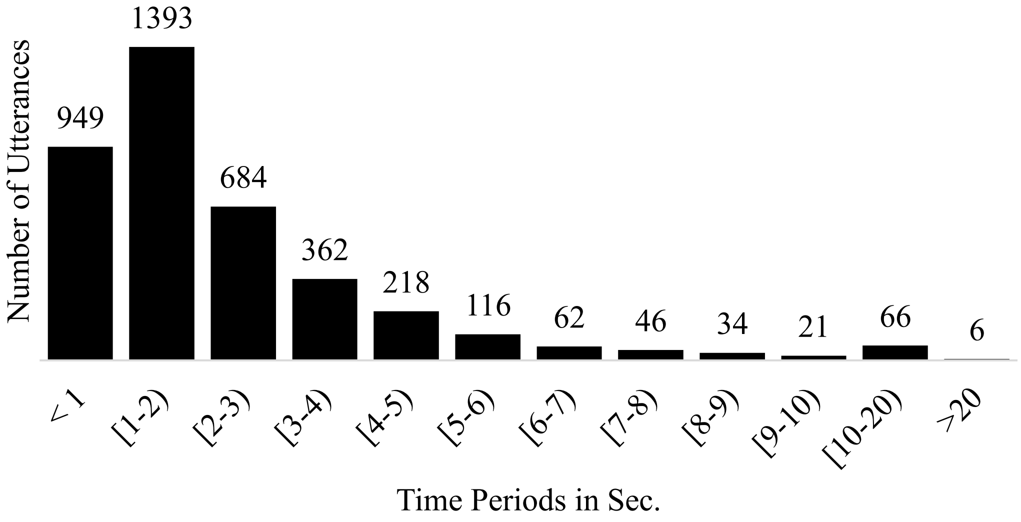

Figure 2.3: Utterance-level audio length distribution.

We examine two types of audio spectrogram feature: Mel-frequency cepstral coefficients (MFCC) and Mel-frequency spectral coefficients (MFSC). MFCC feature extraction has been successfully applied in speech recognition [17] and audio classification [18]. However, MFCC includes the discrete cosine transform (DCT), which can flaw the locality especially for the convolutional neural network (CNN) [19]. Therefore, several audio classification types of research use MFSC instead [20]. Furthermore, we inspect the effect of adding the dynamic features: first and second temporal derivative (delta and the delta-delta coefficients respectively). Adding the dynamic features coefficients can increase the accuracy and the robustness of speech recognition [21]. Table 2.5 shows the evaluation results for both feature types and their derivatives. Results show that MFSC features dominance over the MFCC with and without their derivative. The reason for that is the locality issue produced by DCT of MFCC transform that we discussed above. Also, we notice that adding the derivatives to the MFCC feature type increases the accuracy slightly. While adding the derivatives for the MFSC decades the accuracy. Therefore, we conclude that static MFSC is the best feature type for our dataset and architecture, which we will consider it for the next experiments.

As mentioned in section 2.3 Dataset Collection and Characteristics, our audio data is recorded using two microphones. Consequently, each audio recording includes two channels. We used the two channels to design five different configurations (Table 2.6). In

Feature type		**Accuracy**	**Average F1-score**
MFCC	Static	26.0	0.162
	Dynamic (Δ, ΔΔ)	27.7	0.200
MFSC	Static	**30.8**	**0.231**
	Dynamic (Δ, ΔΔ)	30.0	0.210

Table 2.5: The accuracy and average F1-Score for different features types.

Input Configuration	Number of Samples	Input Dimension
CH1 only	3557	(60, 40, 10)
CH2 only	3557	(60, 40, 10)
Unite CH1 with CH2	3557×**2**	(60, 40, 10)
(CH1 + CH2)/2	3557	(60, 40, 10)
Combine CH1 & CH2	3557	(60, 40, **20**)

Table 2.6: Input channel configuration.

the first and second setups, we include either channel one or two separately. In the third setup, we double the dataset by feeding both channels as a distinct sample. In the fourth and fifth setups, we use both channels together either by sum and average them into a single channel or feed them both as a two-channels. Table 2.7 shows the accuracy and the average F1-score for each one of these five input setups. From the results, we notice that the last two setups, when the two channels are combined, the accuracy is better than the first three setups when the input is one channel only. The reason is that labels transcribed using both channels, so when we omit one of input channels some utterances may have wrong labels, and consequently the neural network fails to predict the activity on the evaluation dataset. While combining the two-channels have higher accuracy. However, average the two channels have slightly lower accuracy than including both channels. Thus, in the next evaluation experiments, we consider the configuration that using the static MFSC feature type and feeding the network with both channels.

Input channels	Accuracy	Average F1-score
CH1 only	22.2	0.106
CH2 only	22.9	0.121
Unite CH1 with CH2	22.6	0.115
(CH1 + CH2)/2	30.2	0.217
Combine CH1 & CH2	**30.8**	**0.231**

Table 2.7: The accuracy and average F1-score for different input channels configurations.

2.5 Model Architectures

We consider the speech-based activity recognition as a multi-class classification problem. In this section, we first present a modified VGG [9] network for the audio branch, which we used to evaluate various configurations discussed in section 2.3. Then, we introduce the new architecture that fuses the output of the proposed audio network with the keyword network to predict the activities.

2.5.1 The Audio Network

Previous neural network architectures designed for image processing have been adjusted successfully to work on audio processing [20], [22] such as VGG [9], ResNet [23], and DenseNet [24]. The VGG network shows a better performance compared with other architectures for audio classification applications [20], [25]. Because deeper CNN networks are easily overfitting on small size datasets. We adapted the VGG network based on the trauma dataset (Figure 2.4). Our modification includes adding a batch normalization [26] to the convolutional neural networks (CNN) to speed up the training operation and assist the regularization. We also use the dropout [27] and Gaussian noise to prevent overfitting and increase generalization. For the activation function, we use rectified linear units (ReLUs) and the last classification layer includes the global average pooling followed by a softmax activation function to calculate the prediction probabilities.

2.5.2 Keyword and Fusion Networks

As shown in Figure 2.5, we designed an architecture that consists of a keyword network, an audio network, and a fusion network. We used a fully connected network (FCN) layer with the ReLU activation function to generate the keyword feature representation. We empirically evaluated different sizes and the number of layers to find

Input 60 × 40 × 20

5 × 5 CNN(128) + BN + ReLU
3 × 3 CNN(128) + BN + ReLU
2 × 2 Max-Pooling
Gaussian-Noise(1.0)

3 × 3 CNN(256) + BN + ReLU
3 × 3 CNN(256) + BN + ReLU
2 × 2 Max-Pooling
Gaussian-Noise(0.75)

3 × 3 CNN(512) + BN + ReLU
Dropout(0.3)
3 × 3 CNN(512) + BN + ReLU
Dropout(0.3)
3 × 3 CNN(512) + BN + ReLU
Dropout(0.3)
3 × 3 CNN(512) + BN + ReLU
Dropout(0.3)
2 × 2 Max-Pooling
Gaussian-Noise(0.75)

3 × 3 CNN(1024) + BN + ReLU
Dropout(0.5)
1 × 1 CNN(1024) + BN + ReLU
Dropout(0.5)

1 × 1 CNN(6) + BN + ReLU
Gaussian-Noise(1.0)
Global-Average-Pooling

6-way Softmax

Figure 2.4: Our audio network architecture. BN: Batch Normalization, ReLU: Rectified Linear Unit.

the optimal network configuration (Table 2.8). The results show that using a single FCN layer with size 128 achieved the best performance. Increased number of FCN layers (deeper) or the number of FCN layer units (wider) decreased the performance. The fusion network concatenated the audio network output features (*a*) and the keyword module outputted features (*w*) into one vector (*y*):

$$y = \gamma(concat(\phi(w),\ \psi(a))) \tag{2.1}$$

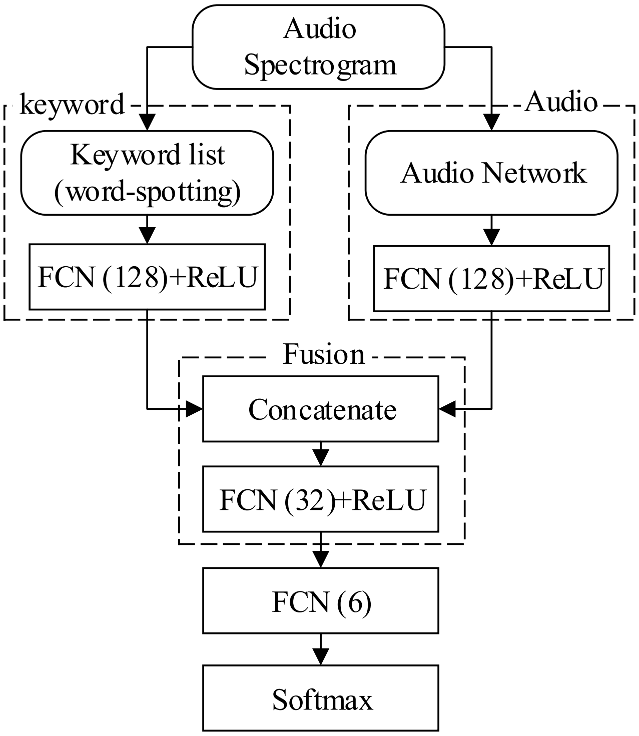

Figure 2.5: Final model archetecture after adding the keyword features. FCN: Fully connected Network, ReLU: Rectified Linear Unit.

Where ϕ, ψ and γ are the fully connected networks layer (FCN), and y is the output of the fusion, which includes another FCN and ReLU activation function to generate the high-level feature representation for the final softmax layer classification (Figure 2.5). We compared different fusion methods such as attention, but it did not perform well due to audio and keyword miss-alignment issues. This issue will be addressed in our future work.

Audio + Keyword	**Accuracy %**	**Average F1-Score**
(1-layer, 64)	44.9	0.412
Deeper (2-layers, 64)	44.9	0.409
Deeper (2-layers, 128)	44.8	0.409
Wider (1layer, 256)	44.7	0.409
(1-layer, 128)	**45.4**	**0.415**

Table 2.8: Results comparison between different keyword and fusion modules layer structure.

2.6 Experiments Setup and Results

We trained and evaluated the proposed model on the trauma resuscitation dataset. We used the utterances from the five most frequent activities and the total number of utterances was 3549. The dataset was randomly shuffled and split into 80% and 20% as a training set and a testing set, respectively. Each sample was considered independently, which contains an utterance-level audio stream, the related one keyword, and the correspondence activity label assigned by the trauma experts. Due to the small data size, we applied the fivefold cross-validation. We trained all networks together as end-to-end using the early fusion approach. We use Adam [28] optimization with 0.001 as the learning rate and categorical cross-entropy loss function. Each experiment took about two hours. We implemented all the experiments using Keras API of the TensorFlow library [29] with two Nvidia GTX 1080 GPUs.

Figure 2.6 shows the hypothetical architectures diagrams for speech-based activity recognition. The first design (Figure 2.6 (a)) integrates an automatic speech recognition

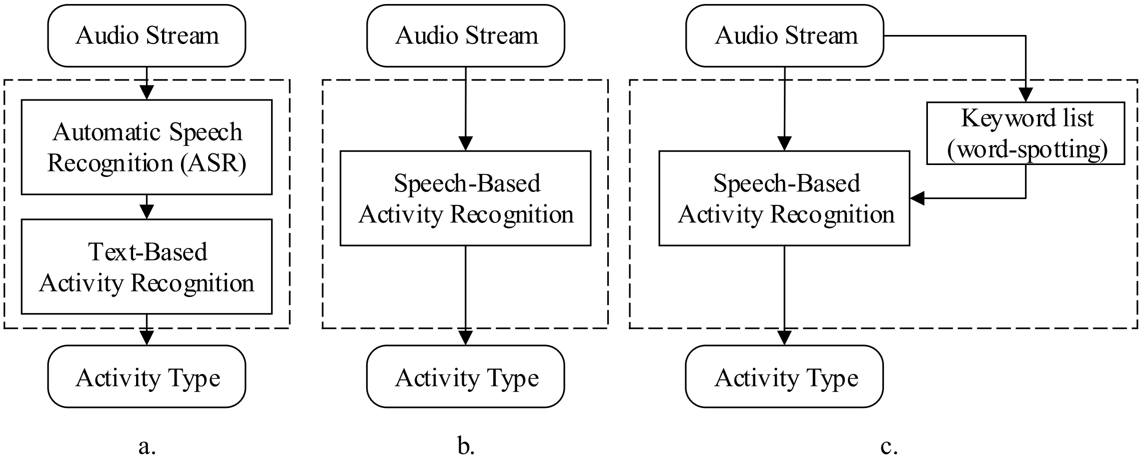
Audio Stream
Automatic Speech Recognition (ASR)
Text-Based Activity Recognition
Activity Type
a.
Audio Stream
Speech-Based Activity Recognition
Activity Type
b.
Audio Stream
Keyword list (word-spotting)
Speech-Based Activity Recognition
Activity Type
c.

Figure 2.6: Speech-based activity recognition proposed architectures. a. an activity recognition that uses the predicted transcripts from automatic speech recognition. b. activity recognition that predict the activity type directly from the audio. c. same as in (b) with an additional one keyword input comes from word-spotting.

(ASR) module with a text-based activity recognition (TAR) module. The overall performance of this design highly depends on both modules. Although the previous TAR model [4] achieves 69.1% accuracy and 0.67 average F1-Score on a resuscitation dataset. Their model used the human transcripts and assumed the ASR system can achieve human parity. Unfortunately, our ASR experiment results show a high word error rate (WER) on the resuscitation dataset using two different architectures: attention-based seq2seq [5] and N-gram [6], which achieved 100.3 and 75.8 WER, respectively. We believe the bad audio quality caused by the distant-talking, trauma noise, fast speaking rate, and concurrent speakers damaged the overall activity recognition performance, which makes the model infeasible. The second architecture predicts the activity type directly from the audio (Figure 2.6 (b)). Our evaluation result shows that the model achieves 30.8% accuracy and 0.231 in average F1-Score. Compared with the above two architectures, the proposed model (shown in Figure 2.6 (c)) achieves 45.4% accuracy and 0.415 in F1-score, which outperforms the previous approaches that using the audio directly or recognizing the full utterance transcript. The comparison result demonstrates the effective and efficiency of using the keyword as an additional feature to the speech-based activity recognition architecture.

We further compare our audio network with other state-of-the-art classification architectures implementations such as VGG16-19 [9], DenseNet [24], ResNet [23] and NASNetMobile [30] (Table 2.9). The result shows that our audio network outperforms others in terms of accuracy and the average F1-score by a range of 1.3% – 8.9% and 0.02 – 0.129, respectively. This is because the general deep architectures usually suffer from overfitting when applied to audio processing [20].

Classification Models	**Accuracy %**	**Average F1-score**
NASNetMobile [24]	21.9	0.102
VGG19 [14]	27.7	0.182
DensNet [17]	28.2	0.190
ResNet [16]	28.0	0.196
VGG16 [14]	29.5	0.211
Our Network	**30.8**	**0.231**

Table 2.9: Results comparison between our network and other classification models.

We made a quantitative analysis by comparing the performance of three different models using different inputs: audio-only, keyword only, and both audio and keyword (Table 2.10). The result shows that using both audio and keyword features outperforms using audio-only and keyword only, which proves our hypothesis that the keyword can boost the performance of the audio-only model, but not enough to replace it.

Table 2.11 shows the F1-Score for each activity by different modalities. We notice that the audio network has better performance on *Extremity* and *Back* activity than *GCS*, *Face,* and *Circulation Control*. Different factors can cause these variations: imbalanced dataset and the noise level. As seen in Table 2.2, the number of utterances that include each activity decreased by 100 from *Extremity* to *Circulation Control* exhibits an unequal distribution between the activities. This imbalance pushes the neural network classifiers to get biased towards the high-frequent activities more than low-frequent activities. For the noise level, Fig. 2.1 shows *GCS* and *Face* have relatively higher noise than other activities, which may impact the prediction performance. The third column of Table 2.11 shows the F1-Scores

Modality Type	**Accuracy %**	**Average F1-score**
Audio only	30.8	0.231
Keyword only	38.3	0.344
Audio + Keyword	**45.4**	**0.415**

Table 2.10: Results comparison between keyword and audio models.

Activity	F1-score		
	Audio	**Keyword**	**Audio + Keyword**
Extremity	**0.366**	**0.532**	**0.524**
Back	**0.448**	**0.375**	**0.582**
GCS Calculation	0.124	0.314	0.313
Face	0.054	**0.389**	**0.385**
Circulation Control	0.045	0.242	0.255
OTHER	**0.351**	0.212	**0.429**

Table 2.11: The F1-Score for each activity for different modalities.

for each activity for the final model that fuses both the audio stream and the keyword. We can see that the scores are boosted almost for all activities.

2.7 Summary

In this chapter, we introduce a new model for the speech-based activity recognition and an empirical assessment on a trauma resuscitation dataset. In our design, we extend the input features to integrate a keyword, a one-word from the most frequent words list that included in the utterance, to the audio stream. The new structure shows a substantial increment in the accuracy and the average F1-score 3.6% and 0.184 respectively over the audio network alone. Due to the high word error rate of the ASR output caused by the fast speaking rate, concurrent speakers, and high noise in trauma resuscitation, our approach relies on an additional one keyword instead of the entire ASR generated utterance is more efficient. Also, we examine the trauma resuscitation audio constraints such as activity recurring, noise level, and most frequent words. In the evaluation results, we found out that the imbalance of the activities in the trauma resuscitation, as well as the noise, reduced the audio network accuracy. Also, we explore audio stream preprocessing factors such as audio channels setup and features type. We found that the static MFSC feature and stereo channel configuration has the best performance. We introduce a new audio network based on the

VGG model and provide an evaluation comparison with various classification architectures. Although our model has relatively fewer layers concerning other classifiers, it overperforms these models. Introducing the keyword features is promising, but we still need further experiments on integrating the word-spotting models with the current architecture to have a more accurate evaluation. Also, we will examine more architectures for the fusion and keyword modules.

CHAPTER 3

KEYWORD ANALYSIS AND MULTIMODAL NETWORK TUNING

3.1 Overview

Automatic activity recognition during medical teamwork can help reduce medical errors and miscommunication during patient care. Activity recognition in settings such as trauma resuscitation is challenging for several reasons, including crowded and noisy workspace, concurrent and fast-paced activities. Speech signals include concurrent speech and loud ambient noise and patient discomfort. Current research on activity recognition mostly relies on visual and sensory data resources to predict human activities [1]–[3]. However, these sources do not apply to the activities that produce no distinct visual cues or do not require medical devices. Examples include the activities where the providers rely on the conversation among them. In these cases, verbal communication associated with these activities can serve as a valuable source of data for activity recognition. Recently, research on predicting activities from speech transcripts [7] and audio recordings [4] showed promising results, although the ambient noise and concurrent speech impaired the system performance. Instead of relying on whole transcripts as input text modality, we pursued an alternative approach that uses only one keyword per utterance. These keywords are used together with the speech signal as inputs to our utterance-level activity recognition system.

Our multimodal network consists of three modules: the audio network, keyword network, and the fusion network. For the audio network, we designed and assessed two approaches to extracting the features from the speech signal. The first design employed a convolutional neural network (CNN) using the VGG [9] design. Our second design replaced several frontend CNN layers with two recurrent neural networks (RNN) to track the temporal information in the speech recordings. For the keyword network, we evaluated five configurations for extracting the features from the keyword input obtained from a manually-generated transcript. We compared two fusions approaches early fusion and late fusion. In addition to the modeling, we performed an extensive analysis of our transcript dataset for keyword selection. The transcripts of verbal communication during 86 resuscitations had each utterance labeled with one or more resuscitation activities. We investigated two approaches to create the candidate keyword lists for different activities. The first approach was based on the frequency of occurrence of words in utterances labeled with the given activity. The second approach relied on computing the sensitivity score of each keyword relative to a given activity. Also, we explored using a different number of keywords per activity to find the optimum number. During the final step, a researcher familiar with trauma resuscitation evaluated the candidate keywords based on their perceived relevance to the given medical activity and removed the low-value keywords. This exploration yielded a significant increase in activity-recognition performance.

3.2 Related Work:

Recently, recognizing human activities using different modalities, such as video or still images, has become an active research area. Most of the current approaches rely on visual features or wearable sensors. Visual sensors, such as an RGB camera or depth camera, provide a powerful source of data that captures people’s actions without obstructing their work. Li et al. [14] developed a method to detect the object location from video frames to assist the activity recognition. In [31], a multimodal network was developed to detect the workflow phase during trauma resuscitation process. Vision-base approaches introduce privacy-related issues, and many medical activities do not have a distinct visual appearance. Passive RFID is another common approach relies on tagging certain medical devices and tracking their movement for recognition of the associated activities [2], [12]. The limitations of passive RFID include the limited types of activities that use taggable devices and the need for continuous tagging of disposable objects. Speech and language have been used as additional sources of data for activity recognition. In [8], the authors built a multimodal system to recognize concurrent activities by using multiple data modalities: depth camera video, RFID sensors, and audio recordings. A separate convolutional neural network (CNN) was used to extract features for each modality, and the features were fused using Long Short-Term Memory (LSTM) network in the final decision layer. This research did not quantify the differences between the performance of different modalities. In [4], the authors created a multimodal transformer network to process manually-transcribed spoken language and the audio sound to predict trauma resuscitation activities. Their quantitative analysis showed average accuracy of 36.4% when using only audio. The accuracy increased to 71.8% when using both modalities. The main drawback of this work is that it was relied on manually-generated transcripts, which is not feasible for contemporaneous use of the activity recognition system to detect human errors and provide alerts. Although our system currently also relies on manually extracted keywords, we believe that keyword spotting is a significantly simpler problem than automatic speech recognition, and we are currently working on automatic keyword spotting.

3.3 Dataset Preprocessing

3.3.1 Dataset Collection and Analysis

The dataset was recorded during 86 trauma resuscitations in the emergency room at a pediatric teaching hospital (Children's National Medical Center)) between December 2016 – May 2017. We obtained approvals from the hospital’s institutional review board (IRB) before the study. All data generated during the study were kept confidential and secure following IRB policies and health insurance portability and accountability act (HIPAA). The audio data was recorded using two fixed NTG2 Phantom Powered Condenser shotgun microphones. These microphones pointed in two locations where the key members of a trauma team normally stand. The recordings were manually transcribed, and each utterance transcript was labeled by its related activity by trauma experts. For this research, we choose five activities that had associated most-frequent utterances in the dataset: Extremity (E), Back (B), Glasgow Coma Scale Calculation (GCS), Face (F), and Circulation Control (CC). All other utterances that do not belong to these activity categories are assigned to the “OTHER” category (Table 3.1).

Our input data is the utterance-level audio stream and one text keyword selected from the keyword list. The audio stream is converted into a spectrogram. Spectrogram representation reduces the dimension of the data and provides better information

#	Activity	Code	Utterances
1	Extremity	E	836
2	Back	BK	701
3	GCS Calculation	GCS	610
4	Face	F	514
5	Circulation Control	CC	407
6	Others	OTHERS	4389

Table 3.1: Resuscitation activities with most utterances.

representation [16]. The audio recordings were sampled at 16MHz. We used 40 filter banks for the short-time Fourier transform with a 2048 window, 25% overlap, and Hann window type. Our neural network required a fixed input length and since 99% of the utterance has a length equal or less than 20 seconds, we resized all the utterance lengths to be 20 seconds by zero-padding the beginning and end of each utterance. However, we empirically test several input lengths and 20 seconds showed the best performance. The final feature map length was 600. Following the work [4], we segmented the input feature map into 10 frame sub-maps to avoid processing distant audio frames. The input sample shape for every single channel was (60, 40, 10). We use the Mel-frequency cepstral coefficients (MFCC) without their temporal derivatives (deltas coefficients) because our empirical experiments show that this representation has the best performance.

The keyword input is encoded as a one-dimensional one-hot vector. The keyword one-hot vector size depends on the total number of keywords in the list. Also, the input vector size is incremented by 0.3 to reduce the one-hot hash function collision probability. For example, the vector size of 234 represents a 180 keywords list (30 keywords per activity).

3.3.2 Keyword Labelling

Keyword labeling involved selecting a word from each utterance as a second input, in addition to the audio sound, to improve the activity prediction. We did keyword labeling in two stages: creating the keyword list and specifying the keyword from the utterance.

A keyword list is a group of words chosen for each activity type to increase activity recognition accuracy. To create the optimum keyword list for given activity recognition, we needed to decide on two factors: (1) the keyword selection method, and (2) the number of keywords in the list associated with each activity type. For keyword selection, we experimented with two methods (Table 3.2). The first method selected the most-frequent keywords (MFKW), based on ranking all the words in the dataset by their frequency of appearance across all the transcripts concerning each activity. The second method was based on keyword sensitivity (KWSN), calculated for each word that appeared in all the utterances associated with a given activity type. The keyword sensitivity represents the prediction score of the activity if the keyword was present, computed as:

$$Sensitivity = \frac{A}{C} \tag{3.1}$$

where *A* denotes the number of all utterances in the dataset related to the given activity that contained the given keyword, and *C* denotes the number of all utterances related to the given activity. Table 3.3 shows a sample of keywords and their sensitivity values for six activities. We created the sensitivity-based keywords list for each activity by ranking the keywords based on their sensitivity value. The KWSN columns in Table 3.2 show the ten keywords with the highest sensitivity values for the six activities. Our system described in Section **Error! Reference source not found.** performed significantly better using the keywords selected based on sensitivity, rather than based on word frequency (Figure 3.1).

#	Extremity		Face		Back		GCS Calculation		Circulation Control		OTHER	
	MFKW	**KWSN**	**MFKW**	**KWSN**	**MFKW**	**KWSN**	**MFKW**	**KWSN**	**MFKW**	**KWSN**	**MFKW**	**KWSN**
1	right	right	blood	blood	pain	spine	gcs	gcs	iv	iv	okay	hi
2	left	left	mouth	right	spine	ok	eyes	eye	io	cc	right	ok
3	okay	ok	right	mouth	okay	step	uh	hi	access	io	gonna	go
4	leg	leg	uh	head	yes	pain	squeeze	yes	right	access	get	right
5	lower	go	left	nare	back	ye	okay	eyes	get	go	alright	us
6	knee	okay	nares	ear	step	offs	open	open	left	get	roll	okay
7	upper	hurt	stable	us	hurt	okay	opening	us	okay	right	look	get
8	extremity	lower	head	left	offs	hurt	hand	go	got	ok	left	lateral
9	hurt	upper	clear	abrasion	right	right	verbal	ok	line	left	let	pulse
10	arm	arm	okay	nares	tenderness	back	right	right	gonna	line	take	ear

Table 3.2: Two lists of ten keywords for six activity types (top row). In the MFKW column of each activity, keywords are ranked by their frequency in the activity-related utterances. In the KWSN column, keywords ranked based on the calculated sensitivity value.

Keyword	Extremity	Face	Back	GCS Calculation	Circulation Control	OTHERS
blood	0.0051	0.0248	0.0033	0.1548	0.0179	0.03
pulses	0.0025	0	0	0	0.0026	0.0267
spine	0.0013	0.1314	0	0.002	0	0.0267
pupils	0	0	0.01	0.0041	0	0.0267
collar	0	0.0058	0	0	0	0.025
neck	0	0.0248	0	0.0061	0	0.0217
back	0.0089	0.0934	0.005	0.0122	0.0102	0.0233
abdomen	0	0	0	0.002	0	0.02
arm	0.0786	0.0015	0.0067	0	0.0153	0.01

Table 3.3: Sensitivity values for samples of words for each activity.

The second factor for the keyword list is the effective number of keywords per activity. We determined the number of keywords per activity empirically, by using different lengths of the keyword lists and evaluating the accuracy of the model. The results showed that the best performance was reached using 30 or 40 keywords per activity (Figure 3.2). A larger or smaller number of keywords showed reduced performance. We selected 30 keywords as the optimum list length.

We observed that the keyword lists generated with either MFKW or KWSN contained ineffective words, such as stop-words (e.g., “the,” “is,” “at,” …), or words unrelated to trauma activities. Stop-words are easy to detect and filter out automatically but to remove the unrelated words we needed to manually analyze all the keyword lists. We categorized

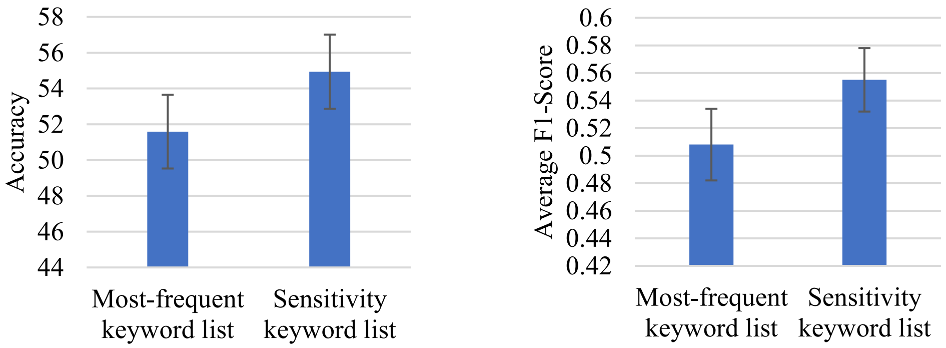

Figure 3.1: Performance comparison of our system for activity recognition using the most-frequent keyword list versus sensitivity-based keyword list.

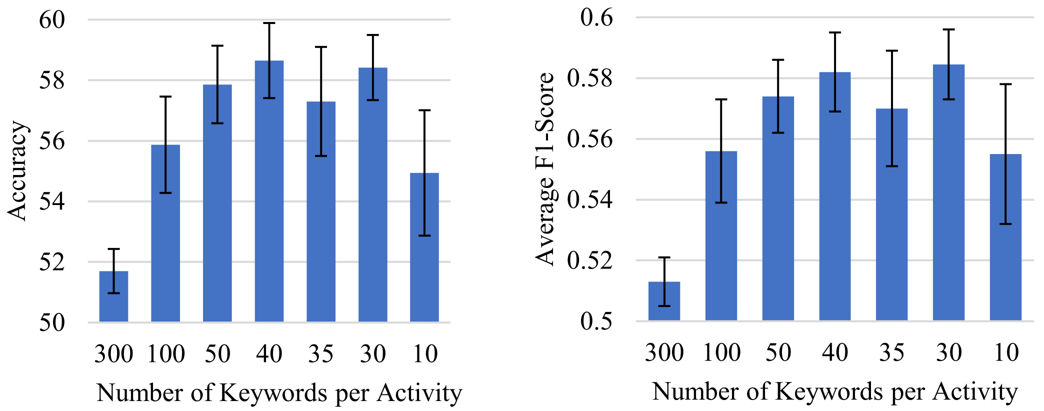

Figure 3.2: Performance comparison results from using different number of keywords per activity.

all the words in dataset transcripts into eleven types based on their relationship to trauma resuscitation activities: acknowledgment, activity, body, body location, equipment, medical term, number, report/request, result, time, and unrelated words. This classification helped us to separate effective words to keep from ineffective words to remove. We found 549 words that did not have a relationship to the trauma resuscitation activity workflow (e.g., “hello,” “something,” and “sorry”). The removal of these unrelated words from the activity-keyword lists increased the system accuracy and the F1-Score by 2.05 and 0.015, respectively, using the 30-keyword list (Figure 3.3).

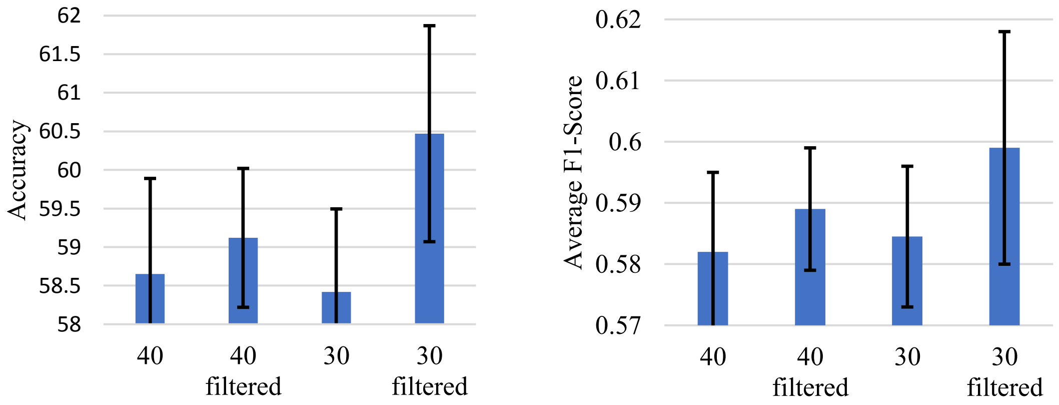

Figure 3.3: Performance comparison between filtered keyword list and non-filtered list.

The second step of keyword labeling is finding the keyword in any given utterance. Based on the generated keyword lists, the utterances can be grouped into three types (Table 3.4):

1. Utterances that do not contain any keyword from the list corresponding to their activity label.
2. Utterances that include a single keyword from the list corresponding to their activity label.
3. Utterances that include more than one keyword from the list corresponding to their activity label.

For the first two types of utterances, we simply selected the keyword "NONE" or the unique keyword, respectively. For the third type of utterances that had multiple keywords from the list, we used the following algorithm (Table 3.5). In the outer loop, we scanned the corresponding keyword list from the highest- to the lowest-ranked keyword. In the inner loop, we scanned the utterance from left to right to search for the current keyword. If we found a matching word, we selected it as the keyword for this utterance. The algorithm stopped when the first keyword was found.

3.4 Model Architecture

We introduced a multimodal neural network for recognition of trauma resuscitation activities from utterance-level audio recording and the selected keyword (Figure 3.4). The model consists of three modules: audio, keyword, and fusion networks. We first

Utterance Type	Number of Utterances
No keywords	1130
Single keyword	1131
Multi keywords	1288
Total	3549

Table 3.4: The number of utterances for each its type.

Utterance	Keyword	Activity
right upper extremity no obvious deformities noted, same with the **left**	**right**	BK
I mean **left** axilla, **left** flank, **left upper extremity**	**left**	GCS

Table 3.5: Two example utterances illustrating the keyword selection algorithm for utterances with multiple candidate keywords, by scanning down the keyword list and scanning the utterance from left to right. Bold words the matching keywords from our list, while the underline denotes the choosing words.

preprocessed the utterance-level speech recording to obtain the spectrogram features. This preprocessing included the algorithms for keyword selection from the corresponding transcript, and for reformatting the keywords into a one-hot representation. Next, audio and keyword networks extracted the features from the corresponding input representations. The fusion network fused the extracted features or decision vectors and made the final prediction from the shared representation.

3.4.1 Audio Network

We experimented with two neural network architectures for acoustic feature extraction (Figure 3.5). Initially, we used a convolutional neural network (CNN), and later replaced

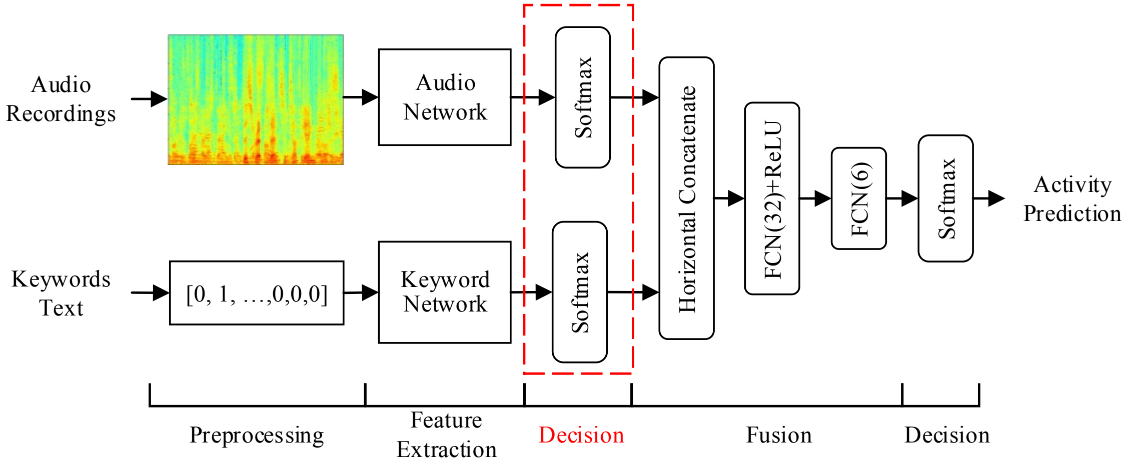

Figure 3.4: Multimodal architecture diagram. The first softmax layers (outlined in a red-dashed box) are included only for late fusion at the decision stage of each network, instead of merging the features directly in the early fusion.

CNN network structure	RNN-CNN network structure
Input 60 × 40 × 20	Input 60 × 40 × 20
5 × 5 CNN(128) + BN + ReLU	Bi-GRU(128) + BN + ReLU
3 × 3 CNN(128) + BN + ReLU	Bi-GRU(128) + BN + ReLU
2 × 2 Max-Pooling	3 CNN(512) + BN + ReLU
Gaussian-Noise(1.0)	Dropout(0.5)
3 × 3 CNN(256) + BN + ReLU	3 CNN(512) + BN + ReLU
3 × 3 CNN(256) + BN + ReLU	1 × 1 CNN(6) + BN + ReLU
2 × 2 Max-Pooling	Dropout(0.5)
Gaussian-Noise(0.75)	Gaussian-Noise(1.0)
3 × 3 CNN(512) + BN + ReLU	Global-Average-Pooling
Dropout(0.3)	
3 × 3 CNN(512) + BN + ReLU	
Dropout(0.3)	
3 × 3 CNN(512) + BN + ReLU	
Dropout(0.3)	
3 × 3 CNN(512) + BN + ReLU	
Dropout(0.3)	
2 × 2 Max-Pooling	
Gaussian-Noise(0.75)	
3 × 3 CNN(1024) + BN + ReLU	
Dropout(0.5)	
1 × 1 CNN(1024) + BN + ReLU	
Dropout(0.5)	
1 × 1 CNN(6) + BN + ReLU	
Gaussian-Noise(1.0)	
Global-Average-Pooling	

Figure 3.5: The structure of the audio network. Our initial CNN model (left) was replaced with an RNN-CNN (right).

the frontend CNN with a recurrent neural network (RNN). The CNN-based model (Figure 3.5, left side) was evolved from the current audio classification research [20], [22], [25]. Our design relied on the VGG network-style [9] because it outperformed other state-of-the-art neural networks on audio classification [20]. The speech signal was represented in a spectrogram format, which converted the time sequence audio into two frequency-time dimensions. This format allowed CNN to process the speech signal similar to images. CNN has a good ability to capture the spatial dependencies between the features from adjacent frames. However, in addition to the frequency domain, the spectrogram represents a time sequence characteristic, which CNN cannot track the sequential information. Instead, RNNs are used to capture the long-range temporal information in sequential data [32]. Therefore, in our second model (Figure 3.5, right side), we replaced the first eight CNN layers with two bi-directional Gated Recurrent Units (Bi-GRUs). The evaluation results showed that RNN-CNN outperforms the CNN model in the accuracy and F1-Score by

1.5% and 0.021, respectively (Figure 3.6). The reason is that the two frontend RNN layers assisted the model to track the temporal dependencies among the sequential speech samples.

3.4.2 Keyword Network

The input of the keyword network is encoded as one-hot representation where the coded position for that spotted keyword is (1) and the rest of them are (0) one word per utterance. We designed and evaluated five different neural networks to determine the best keyword feature extraction (Figure 3.7). We used different types of layers to build our networks: two models used fully connected networks (FCN); two models used CNN, and the last one used an RNN. The difference between the two versions of FCN and the two of CNN was in the network depth (Figure 3.7, first four columns). The evaluation results showed that, in general, fully connected network networks outperformed both RNN and CNN networks (Figure 3.8). The deeper FCN II achieved the highest accuracy and F1-Score. The reason is that the input information had no sequential or spatial dependency to be tracked, which limited the performance of RNN and CNN networks.

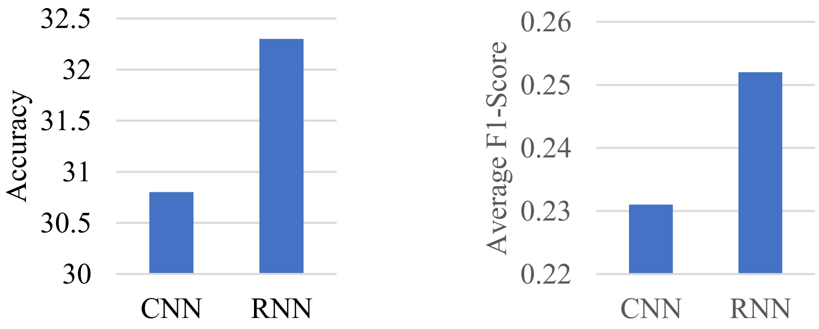

Figure 3.6: Comparison between CNN and RNN audio networks.

FCN I	**FCN II**	**CNN I**	**CNN II**	**RNN**
Input (180, 1)	Input (180, 1)	Input (180, 1)	Input (180, 1)	Input (180, 1)
FCN (128)+ReLU	FCN (1024)+ReLU	3×3CNN(256)+BN+ReLU	3×3CNN(128)+BN+ReLU	Bi-GRU(128)
FCN (64)+ReLU	FCN (512)+ReLU	3×3CNN(128)+BN+ReLU	3×3CNN(128)+BN+ReLU	Bi-GRU(128)
FCN (32)+ReLU	Dropout(0.3)	3×3 CNN(64)+BN+ReLU	Gaussian-Noise(1.0)	Faltter ()
FCN (6)	FCN (256)+ReLU	Global-Average-Pooling	3×3CNN(256)+BN+ReLU	FCN(64)+ReLU
	Dropout(0.3)	FCN (32)+ReLU	3×3CNN(256)+BN+ReLU	FCN(64)+ReLU
	FCN (128)+ReLU	FCN (6)	Gaussian-Noise(0.75)	FCN(64)+ReLU
	FCN (64)+ReLU		3×3CNN(512)+BN+ReLU	FCN(6)
	FCN (32)+ReLU		Dropout(0.3)	
	FCN (6)		3×3CNN(1024)+BN+ReLU	
			Dropout(0.3)	
			3×3CNN(1024)+BN+ReLU	
			Dropout(0.3)	
			3×3 CNN(256)+BN+ReLU	
			Gaussian-Noise(0.75)	
			3×3 CNN(128)+BN+ReLU	
			Dropout(0.5)	
			1× CNN(64)+BN+ReLU	
			Dropout(0.5)	
			1×1 CNN(6)+BN	
			Gaussian-Noise(0.3)	
			Global-Average-Pooling	
			FCN (6)	

Figure 3.7: Keyword networks: FCN I = fully connected network shallow model; FCN II = fully connected network deep model; CNN I = convolutional neural network shallow model; CNN II = convolutional neural network deep model; RNN = recurrent neural network model.

3.4.3 Fusion Network

In our architecture, the fusion stage is to project different modalities (audio, keyword) into a shared representation space for final decision-making. Fusion had been applied in several applications that used multiple modalities. Multimodal fusion provides additional

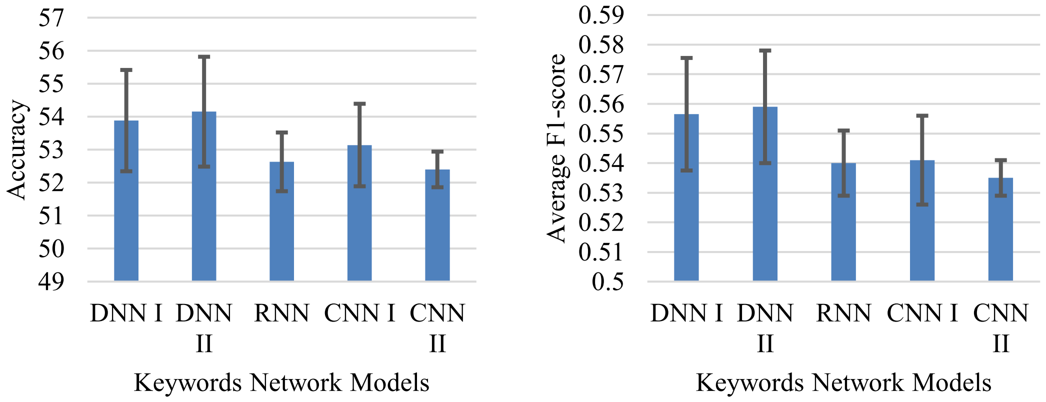

Figure 3.8: Comparison between five different keyword networks.

information that increases decision-making accuracy [31]. According to previous research, there are two widely used fusion: early fusion and late fusion.

Early fusion (or feature-level fusion) merges the features extracted from various modalities such as visual features, text features, and audio features, into an integrated feature vector before feeding it into the classifier. The correlation between different features at an early phase can increase task performance. However, weak synchronization between temporal features in different modalities can degrade the analysis results. Figure 3.4 shows the multimodal design for both early and late fusion approaches. If we exclude the dotted decision box then the diagram characterizes the early fusion approach, where the horizontal concatenation layer used to include both features that come from audio and keyword modules.

Unlike early fusion, late fusion (or decision-level fusion) merges the decisions that produced from different modality networks and then generates cross-modal interactions for prediction. In late fusion, processing the fusion vector is easier than in the early fusion, because the merged decisions have a similar format. However, late fusion requires more computation and can lose the correlation between different features. Including the dotted decision box, Figure 3.4 characterizes multimodal design using the late fusion approach, where the horizontal concatenation layer used to include decision outcomes from audio and keyword modules. In this approach, we first trained the audio and keyword networks. We then used these trained models to predict both the training and the testing datasets. The predicted results of the training dataset of both networks were combined using the horizontal concatenation layer to train the fusion network. Finally, the predicted outcome from the testing dataset was fed to the fusion network, and its predicted output was used to

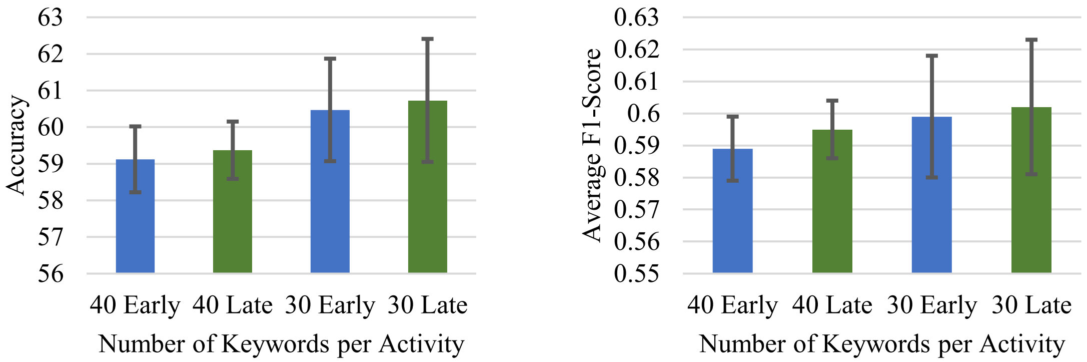

Figure 3.9: Comparison between early and late fusion methods.

evaluate the model performance. We found that the late fusion approach outperformed the early fusion approach (Figure 3.9). Late fusion outperformed early fusion by 0.26 and 0.005 for accuracy and F1-Score, respectively. The reason is that there is no synchronization between the temporal speech features and the non-temporal keywords features. Therefore, in our case, merging the decisions is better than merging the features.

3.5 Experiments Setup and Results

3.5.1 Experiments Setup

We trained and evaluated the proposed model on the trauma resuscitation dataset containing 86 transcripts. We used the utterances from the five most frequent activities and part of the “OTHERS”, for which the total number of utterances is 3549 (Table 3.4). Due to the small dataset size, we applied five-fold cross-validation. We also randomly shuffled the dataset and split it into 80% and 20% as a training set and a testing set, respectively. Each data sample contained an utterance-level audio stream, a single selected keyword, and the manually assigned ground-truth activity label. We used Adam optimization [28] with 0.001 as the learning rate and the categorical cross-entropy loss function. We implemented all the experiments using Keras API of the TensorFlow library [29] with two Nvidia GTX 1080 GPUs. The runtime for each experiment was about two hours.

3.5.2 Results and Discussion

We first compare the model performance for different types of utterances, grouped by the number of keywords (Figure 3.10). This comparison illustrates the importance of the keyword module on performance because the predication relies only on the audio module when the utterance had no keyword. As seen in **Error! Reference source not found.**, the accuracy and the F1-Score for the utterances that contain keyword have almost double the performance for the utterances without any keywords.

We also provide the confusion matrix of the average of the five-fold training-testing experiments. Values in the matrix cells represent the number of observations for each activity (Figure 3.11). The y-axis represents the true activities, while the x-axis represents the predicted activities. As seen, the true positive values, when the predicted activity is equal to the ground truth (the dark diagonal) signify most of the outcomes. While, in general, the false observation whether it is positive or negative have relatively small values. We notice that the "OTHERS" activity has higher false-positive values concerning other

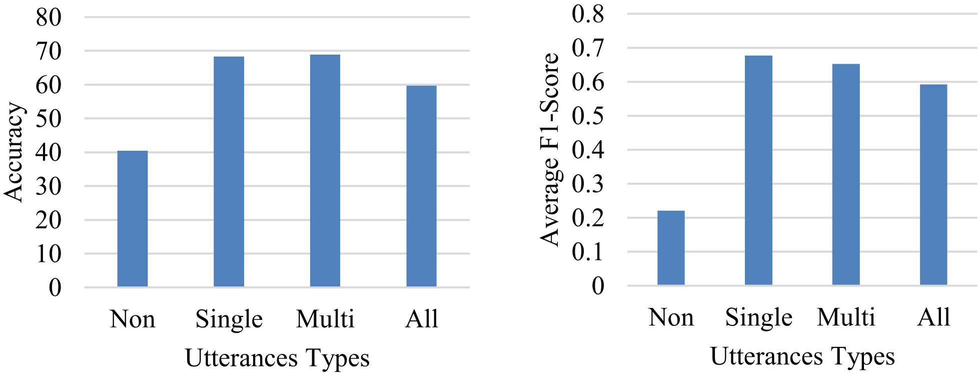

Figure 3.10: Performance comparison between utterances types.

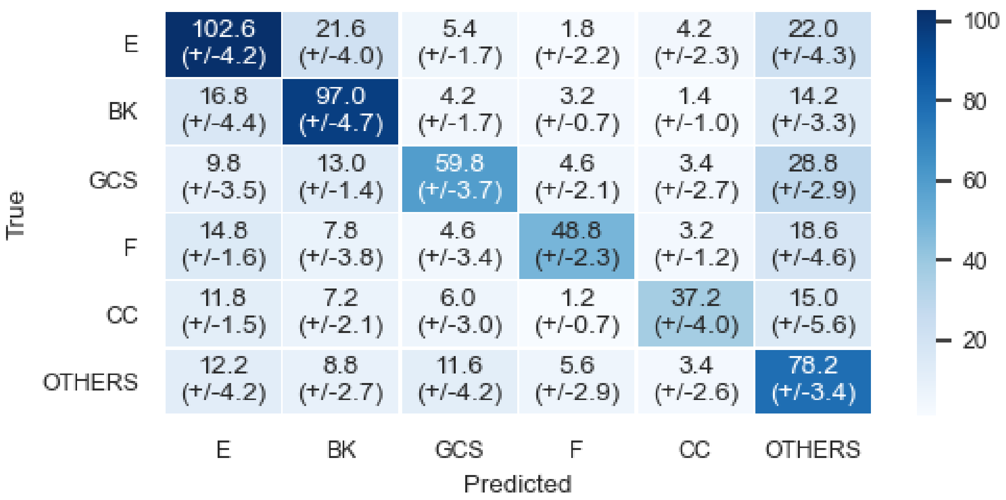

Figure 3.11: The confusion matrix for the activities. X-axis represent the true activities labels, while the y-axis represent the predicted activities. Each value represents the average number of observations for the five-fold training/testing sets, whereas the values inside the parentheses represent the standard deviation.

activities and that is because its utterances belong to several activities, which is expected to be harder for the model to distinguish.

We further show the F1-Score for each activity to disclose the performance for each activity individually (Figure 3.12). As seen, the F1-Score of the Back (BK) and Extremity (E) is the highest at 0.664% and 0.63, respectively. While the least F1-Score is for the “OTHERS” activity with 0.527. Finally, we demonstrate the performance improvement

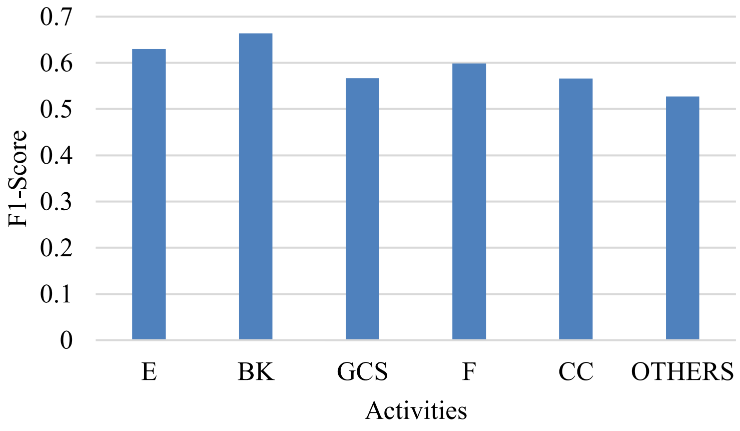

Figure 3.12: Comparison between activities F1-Score.

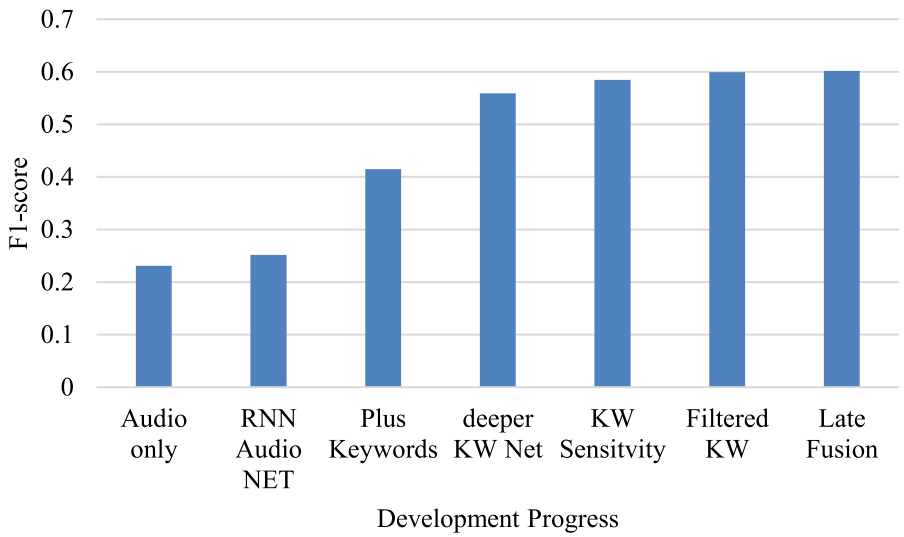

Figure 3.13: The F1-Score improvement over the architecture evolvement.

over our model evolvement (Figure 3.13 and Figure 3.14). The comparison includes seven different model designs that developed cumulatively from the baseline using the audio modality only with the CNN network with 30.8 and 0.231 accuracy and F1-Score, respectively. The final best model involves using keywords in addition to the audio with the best perform keyword list, modality network, and fusion approach. This model has 60.73% and 0.602 accuracy and F1-Score, respectively that is an increment over the baseline model by 29.93% and 0.371 accuracy and F1-Score, respectively.

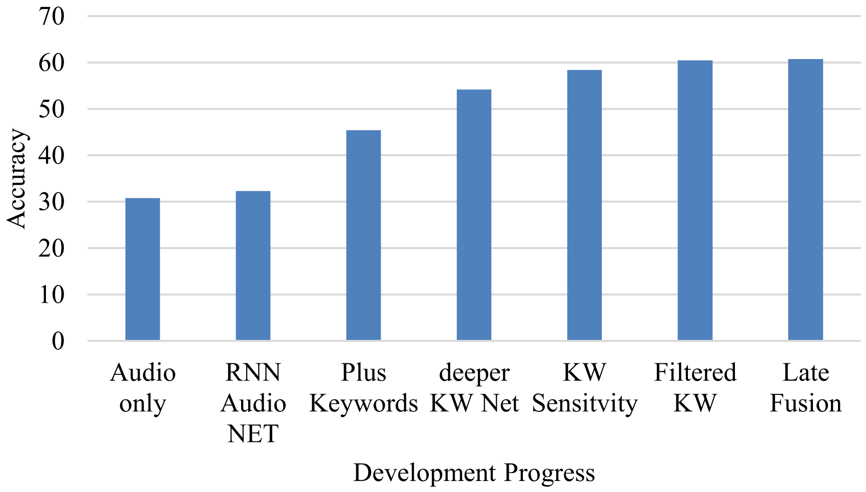

Figure 3.14: The accuracy improvement over the architecture evolvement.

3.6 Summary

We introduced a multimodal neural network that uses the audio recording and the related text keyword to predict the trauma activity associated with each utterance. Our model includes three neural-network modules: audio, keyword, and fusion networks, to process the audio signal spectrogram representation, the one-hot keyword encoding, and the combined output of the audio and keyword networks, respectively. We evaluated different network types for the network modules and keyword selection approaches. A combination of RNN-CNN, deep FCN, and late fusion approaches showed the best performance for the audio, keyword, and fusion networks, respectively. Using the keyword list based on the sensitivity value, followed by manual filtering of keywords, showed the best performance. The final model that evolved by network tuning and the analysis of keyword selection increased the activity prediction performance by 29.93 and 0.371 in terms of the accuracy and F1-Score, respectively.

CHAPTER 4

SPEECH ENHANCEMENT ARCHITECTURE[1]

4.1 Overview

Speech enhancement has important applications in voice communication, hearing aids, and automatic speech recognition. Speech enhancement removes background noise from noisy speech signals, increasing speech quality, and intelligibility [33], [34]. Early research used non-trainable statistical approaches on spectrograms, such as spectral subtraction [35], Wiener filter [36], statistical model-based methods [37], the subspace method [38], minimum mean-square error estimator, and optimally-modified log-spectral amplitude [39], [40]. These methods showed limited performance on speech with non-stationary noise, which is common in real-life environments. Non-negative matrix factorization has later been widely used for speech separation and enhancement [41], [42].

Recently, deep neural networks have been employed to overcome the non-stationary condition and have improved speech quality and intelligibility. Early models used mapping-based methods, where the enhanced signal is directly predicted from the noisy one. Several such deep learning models have been developed, including denoising autoencoders [43] (using fully-connected layers), recurrent neural networks (RNN) [44], and convolutional neural networks (CNN). Later, a masking-based method was introduced to enhance the signal by applying the noisy signal to the predicted mask [45]–[48].

[1] This chapter is based on a published paper: J. Abdulbaqi, Y. Gu, S. Chen and I. Marsic, "Residual Recurrent Neural Network for Speech Enhancement," in ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020.

Most of these methods use time-frequency (T-F) spectrogram features instead of time-domain waveform since T-F has a reduced resolution. Spectrogram features, however, have certain limitations. First, the pre- and post-processing operations such as discrete Fourier transform and its inverse are computationally expensive and cause artifacts in the output signal [33], [34]. Second, these approaches usually only estimate the magnitude and use the noisy phase to produce the enhanced speech. Research has shown that the phase can enhance speech quality [49]. Recent research has considered predicting the phase and the magnitude at the cost of model complexity, such as adding a special model for phase component [50].

Recently, several studies proposed overcoming previous limitations by working directly on the waveform. Fu et al. [51] compared fully-convolutional networks with fully-connected networks. Pascual et al. [52] implemented a generative adversarial network for speech enhancement (SEGAN), using strided convolutions, residual connections, and an encoder-decoder architecture. Later, a text-to-speech model called Wavenet [53] directly synthesized raw waveforms. Qian et al. [54] and Rethage et al. [55] presented a modified version of WaveNet for speech denoising. The former integrated a Bayesian framework WaveNet, while the latter used a non-causal dilated convolution with residual connections. Germain et al. [56] presented dilated convolutions combined with a feature loss network. Stroller et al. [57] adapted the U-Net [58] model for source separation using dilated convolutions and linear interpolation instead of transposed convolution for upsampling. All these methods used convolutional neural networks due to their ability to capture the samples' dependencies better than fully connected networks. Because waveform is a sequential data type, it requires a temporal context as well. Recurrent neural networks are known to capture the long-range temporal sequence information [32] and are used in many sequential applications such as speech recognition, neural machine translation, and spectrogram-based speech enhancement. To our knowledge, only [59] and [60] have applied RNN to process waveform signals. The first one used RNN to denoise a non-speech waveform, while the latter used RNN for speech bandwidth extension, but no one has used it for waveform-based speech enhancement. The reason is that the high resolution of waveforms requires more expensive, deeper, and wider networks. It is difficult to build a deep RNN because of saturating activation function, which causes gradient decay over layers. Also, we found empirically that RNNs sufficiently wide to process the high-resolution waveforms exceeded the available memory capacity. Therefore, we introduce our residual hourglass recurrent neural network for waveform-based single-channel speech enhancement. Our model overcomes the RNN limitations by introducing two techniques. First, the network architecture has an hourglass shape; the layers in the lower pyramid reduce the number of time-steps and increases the number of units (width), while the upper pyramid does the reverse. This architecture allows the RNN to handle high-resolution waveform features without memory overflow. Second, using residual connections between the same-shaped layers from the lower pyramid to the upper one prevents gradient decay over layers and improves the model generalization.

4.2 Model Architecture

Our model includes seven GRU layers with two residual connections. The first six layers are bidirectional and the last one is a single GRU (Figure 4.1). The goal of our speech enhancement network is to learn non-linear relationships, so that noisy speech $x(t)$ can be translated into clean speech $y(t)$:

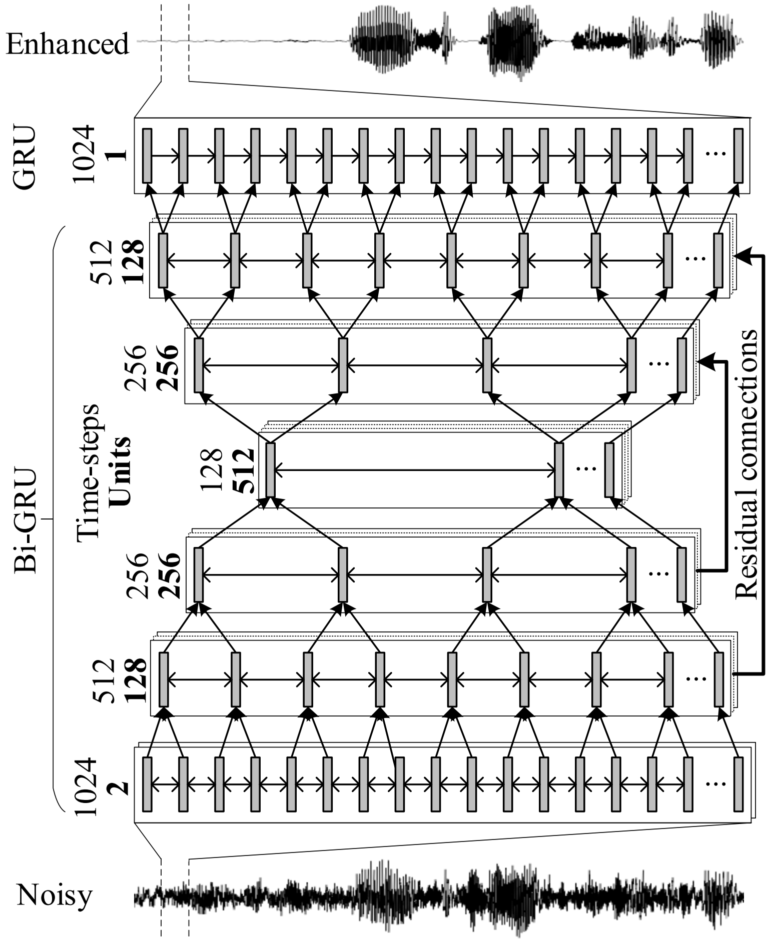

Figure 4.1: Our proposed RNN architecture. Seven stacked RNN layers with the numbers on the left representing the number of time steps and the number of units in each layer. Wider layers have fewer units and vice versa. The two bold arrows on the right represent the residual connections.

$$y(t) = f(x(t)) \tag{4.1}$$

The input vector $X = (x_1, \dots, x_T)$ represents a T-seconds wide segment from a noisy audio waveform signal.

RNNs can efficiently realize temporal features in sequential data, so they have been used widely to process speech data either for speech recognition or enhancement. We chose gated recurrent units (GRU) instead of long short-term memory units (LSTM) or vanilla RNN. Both GRU and LSTM outperform vanilla RNN [28], but GRUs have a simpler structure and train faster than LSTMs. Besides, we chose bidirectional RNNs since in

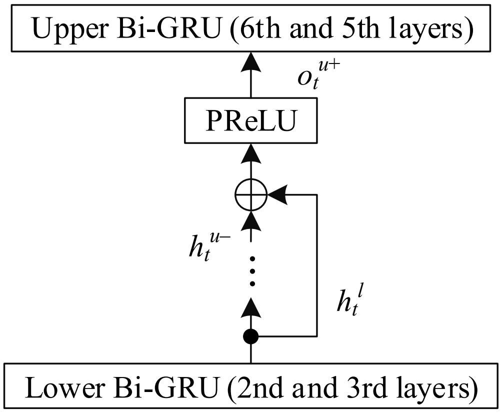
Upper Bi-GRU (6th and 5th layers)
o_t^{u+}
PReLU
h_t^{u-}
h_t^l
Lower Bi-GRU (2nd and 3rd layers)

Figure 4.2: A high-level view highlighting the residual connections in our proposed model from Figure 4.1.

speech enhancement each predicted sample can depend on the future as well as past noisy samples. The stacked GRU increases the capacity of the network by sharing the hidden states not only from the same layer but also from the lower layers as well. The stacked bidirectional RNNs share their hidden states so that the hidden state (h_t^l) of a bi-GRU unit in layer *l* at time *t* is obtained by concatenating its forward ($\overrightarrow{h_t^l}$) and backward ($\overleftarrow{h_t^l}$) hidden states, which depend on the lower layer *l*–1 at time *t* and this layer at time *t*–1:

$$\overrightarrow{h_t^l} = \overrightarrow{GRU}\left(\overrightarrow{h_t^{l-1}}, \overrightarrow{h_{t-1}^l}\right) \tag{4.2}$$

$$\overleftarrow{h_t^l} = \overleftarrow{GRU}(\overleftarrow{h_t^{l-1}}, \overleftarrow{h_{t-1}^l}) \tag{4.3}$$

$$h_t^l = conc(\overrightarrow{h_t^l}, \overleftarrow{h_t^l}) \tag{4.4}$$

The two pyramids of our hourglass architecture keep the number of trainable parameters within the memory constraints. the bottom pyramid decreases the number of time steps while increasing the number of GRU units per layer, and the top pyramid does the reverse. this approach allows for deeper networks. we did not use upsampling techniques, such as linear interpolation, because the information can be lost. instead, we reshape the RNN output to the desired fewer time steps. reshaping the layer output to decrease and increase the time steps prevents losing data and allows the RNN to have a sufficient size of units. however, while stacking RNNs can increase the capacity of the network, deeper RNNs usually have gradient decay issues due to their saturating activation functions. to address this issue, we used residual connections between the lower and upper layers (Figure 4.1 and Figure 4.2). the residual connections facilitate training the deep RNN and provide better generalization by combining the low-level features with the high-level ones in the upper layers. in Figure 4.2, the hidden states of the lower layer ($\boldsymbol{h_t^l}$) and those of the upper layer before the residual connection ($\boldsymbol{h_t^{u-}}$) are combined to produce the residual output:

$$o_t^{u+} = PReLU\left(h_t^l + h_t^{u-}\right) \tag{4.5}$$

$PReLU$ is the parametric rectified linear unit activation function. Finally, we use a single forward GRU to output the enhanced speech with the same size of the input vector:

$$\overrightarrow{h_t^l} = \overrightarrow{GRU}\left(\overrightarrow{h_t^{l-1}}, \overrightarrow{h_{t-1}^l}\right) \tag{4.6}$$

Therefore, the output will be created by combining the hidden states for each input segment:

$$Y = (\overrightarrow{h_1^7}, \dots, \overrightarrow{h_T^7}) \tag{4.7}$$

where Y denotes the enhanced signal output and $\overrightarrow{h_1^7}$ denotes the hidden state of the last (seventh) layer.

4.3 Dataset and Preprocessing

The dataset used for training and evaluating our model has been set up in [61]. We chose this dataset because it is large, has different types of non-stationary noise, and is public so that we can compare our results with other published work. This dataset is an excerpt from the Voice Bank corpus [62] with 28 speakers (14 male and 14 female) of the same accent region (England) and another 56 speakers (28 male and 28 female) of other accent regions (Scotland and United States).

The noisy data used for training are two artificially generated (speech shaped noise and babble) and eight real noise recordings from the Demand database [63]. The noises are from different environments such as kitchens, offices, public spaces, transportation stations, and streets. The training set includes 11,572 utterances with four signal-to-noise (SNR) values: 15 dB, 10 dB, 5 dB, and 0 dB.

The noisy data used for testing include two other speakers of the same corpus from England (a male and a female) and five other noises from the Demand database. The chosen noises include a living room, an office, a bus, and street noise. The testing set includes 824 utterances with four SNR: 17.5 dB, 12.5 dB, 7.5 dB, and 2.5 dB. We downsampled the audio signals to 16kHz, getting a reasonable dataset size for recognizing speech. Our preprocessing included slicing both noisy and clean speech signals into 1024 samples (~64 ms) with a 25% overlap during training and without overlap during the evaluation. We did not use any other preprocessing, such as pre-emphasis.

4.4 Experiment Setup and Results

Our architecture uses seven GRU layers. The first six are bi-directional, while the last one is single-directional to produce the enhanced signal (Figure 4.1). The number of units per layer is 2, 128, 256, 512, 256, 128, and 1; the size of the time steps per layer are 1024, 512, 256, 128, 256, 512, and 1024. Two residual connections link the second and third layers with the sixth and fifth, respectively. The PReLU activation function is used with residual connections, as it does not saturate the negative values compared to Leaky-ReLU and has been shown to improve model fitting [64]. The model has 2 million trainable parameters, which is small concerning Wavenet which has 6.3 million. We use the Xavier normal initializer [65] for the kernel weights, with zero-initialized biases. Xavier's initialization keeps the values of the weights in a reasonable range, preventing the inputs from shrinking or growing more than needed through the layers. It determines the initialization values concerning the number of input and output neurons. The initializer for the recurrent states is a random orthogonal matrix [66], which helps the RNN stabilize by avoiding vanishing or exploding gradients. The stability occurs because the orthogonal matrix has an absolute eigenvalue of one, which avoids the gradients from exploding or vanishing due to repeated matrix multiplication.

We use the log-cosh loss function, a regression loss function that takes on the behavior of squared-loss when the loss is small, and absolute loss when the loss is large; this reduces the influence of wrong predictions. The optimization algorithm used is RMSprop [67], which helps the training of large neural networks on large redundant datasets. Also, Keras [68] documentation recommends using this algorithm with RNN. We trained the model until the validation loss converged with a batch size of 512, using two NVIDIA GTX-1080 GPUs. We used different learning rates during training, starting at 10^{-4} and gradually decreasing to 10^{-8}. The library used to implement this work was Keras with TensorFlow [29] as a backend. The training process took about 20 hours for 50 epochs. To evaluate our model, we computed six objective metrics using an open-source implementation[2,3]:

[2] https://www.crcpress.com/downloads/K14513/K14513_CD_Files.zip

[3] http://ceestaal.nl/stoi.zip

- Segmental signal-to-noise ratio (SSNR) [33]: computed by dividing the clean and enhanced signals into segments and computing the segment energies and SNRs, and then returning the mean segmental SNR (dB). The values range from -10 to 35.
- Perceptual evaluation of speech quality (PESQ) [33]: a more complex metric to capture a wider range of distortions. PESQ is the most common metric to evaluate the speech quality, calculated by comparing the enhanced and clean speech. The values range -0.5 to 4.5.
- Short-time objective intelligibility (STOI) [69]: reflects the improvement in speech intelligibility with a score range from 0 to 1.
- Three objective versions of mean opinion scores (MOSs): CSIG for signal distortion evaluation, CBAK for noise distortion evaluation, and COVL for overall quality evaluation. We used their mathematical representations, and their scores range from 1 to 5 [33].

For all these metrics, higher values mean better performance. Two speech test samples (small segment 50 ms) are illustrated in Figure 4.3. Both samples include non-stationary noise with people talking in background ("cocktail party") and music playing. For each segment, the foreground speaker talks (high frequency) in the first half, while the foreground speaker stops talking (low frequency) in the second half. The enhanced speech signal tracks the clean in both cases, which shows the model's ability to capture the clean speech in all speaker events.

Model	Features type	SSNR	PESQ	STOI	CSIG	CBAK	COVL
No Enhancement (Noisy)	-	1.68	1.97	0.820	3.35	2.44	2.63
SEGAN, 2017 [20]	waveform	7.73	2.16	0.93	3.48	2.94	2.80
CNN-GAN, 2018 [14]	spectrogram	-	2.34	0.93	3.55	2.95	2.92
Wavenet, 2018 [23]	waveform	-	-	-	3.62	3.23	2.98
MMSE-GAN, 2018 [16]	spectrogram	-	2.53	-	3.80	3.12	3.14
DFL, 2019 [24]	waveform	-	-	-	3.86	3.33	3.22
Our model	waveform	**14.71**	**3.20**	**0.98**	**4.37**	**4.02**	**3.82**

Table 4.1: Evaluation results of our proposed model compared with other state-of-the-art research work using six objective metrics on the same dataset [61]. Higher scores are better, and the highest scores are boldfaced.

Table 4.1 shows the metrics score for our model concerning the other architectures. The comparison results with several current architectures such as SEGAN [52], Mask-based GAN model (CNN-GAN) [46], wavenet model for denoising [55], another masking-based GAN model [48] and finally the speech denoising with deep feature losses (DFL) [57]. All these architectures use the same dataset and the metrics that we used to train and evaluate our model. Therefore, their results are taken directly from their above work. In this results, our model decreases the speech degradation (CSIG) by 13.2% and decrease the background noise intrusiveness (CBAK) by 20.7%, and increase the overall signal quality (COVL) by 18.6% concerning the best previous architecture DFL [56]. Also, the speech quality is increased by 26.5% concerning the masking-based GAN model [48].

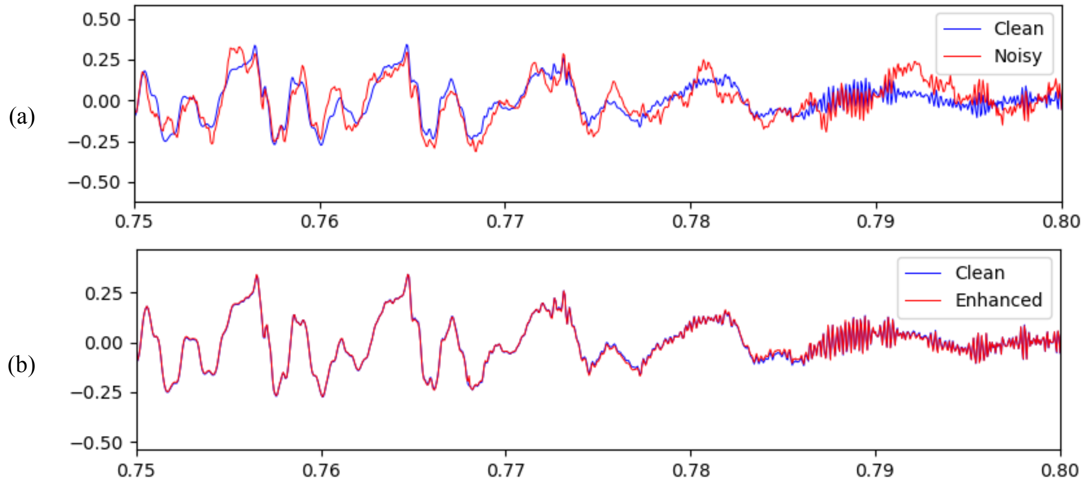

Figure 4.3: An illustration of speech enhancement using our model using speech samples with SNR = 2.5 dB and duration of 50 msec. from the test dataset. (a) The sample number 232_052. The blue lines represent the clean speech and the red lines represent the noisy speech. (b) The corresponding enhanced speech (red line) compared with the clean input speech (blue line).

4.5 Summary

We introduced a novel end-to-end fully recurrent neural network for single-channel speech enhancement. Our recurrent layers are designed in an hourglass shape to reduce the speech signal dimension and assist recognition of the long-term dependencies. The results show that our simple and efficient model outperforms most of the current approaches with more complex architectures. We will evaluate this model with other datasets and apply them to other sequential applications.

CHAPTER 5

CONCLUSION

We introduce a new model for the speech-based activity recognition and an empirical assessment on a trauma resuscitation dataset. In our design, we extend the input features to integrate a keyword, a one-word from the most frequent words list that included in the utterance, to the audio stream. The new structure shows a substantial increment in the accuracy and the average F1-score 3.6% and 0.184 respectively over the audio network alone. Due to the high word error rate of the ASR output caused by the fast speaking rate, concurrent speakers, and high noise in trauma resuscitation, our approach relies on an additional one keyword instead of the entire ASR generated utterance is more efficient. Also, we examine the trauma resuscitation audio constraints such as activity recurring, noise level, and most frequent words. In the evaluation results, we found out that the imbalance of the activities in the trauma resuscitation, as well as the noise, reduced the audio network accuracy. Also, we explore audio stream preprocessing factors such as audio channels setup and features type. We found that the static MFSC feature and stereo channel configuration has the best performance. We introduce a new audio network based on the VGG model and provide an evaluation comparison with various classification architectures. Our model, that has relatively fewer layers concerning other classifiers overperforms them. Introducing the keyword features is promising, but we still need further experiments on integrating the word-spotting models with the current architecture to have a more accurate evaluation. Also, we will examine more architectures for the fusion and keyword modules.

In this work, we present a multimodal neural network that processes the audio recording and their related text keywords to predict the trauma activities. Our model includes three modules: Audio, keyword, and fusion networks to process the audio signal spectrogram representation, the one-hot keyword encoding, and the combined output of the networks, respectively. For the audio network, we design and evaluate two neural networks. The first one is structured with a convolutional neural network. While in the second, we replace the front-end layers by two bi-directional recurrent neural networks to track the temporal dependency of the sequential audio data, which shows a better performance in the evaluation results. For the keyword network, we design and evaluate five networks using different layers types.

We introduced a novel end-to-end fully recurrent neural network for single-channel speech enhancement. Our recurrent layers are designed in an hourglass shape to reduce the speech signal dimension and assist recognition of the long-term dependencies. The results show that our simple and efficient model outperforms most of the current approaches with more complex architectures. We will evaluate this model with other datasets and apply them to other sequential applications.

REFERENCES

[1] I. Chakraborty, A. Elgammal, and R. S. Burd, "Video based activity recognition in trauma resuscitation," *IEEE International Conference on Automatic Face & Gesture Recognition*. pp. 1–8, 2013, doi: 10.1109/FG.2013.6553758.

[2] X. Li, Y. Zhang, I. Marsic, A. Sarcevic, and R. S. Burd, "Deep Learning for RFID-Based Activity Recognition," *International Conference on Embedded Networked Sensor Systems*, vol. 2016. pp. 164–175, 2016, doi: 10.1145/2994551.2994569.

[3] A. Avci, S. Bosch, M. Marin-Perianu, R. Marin-Perianu, and P. J. M. Havinga, "Activity Recognition Using Inertial Sensing for Healthcare, Wellbeing and Sports Applications: A Survey," *ARCS Workshops*. pp. 167–176, 2010.

[4] Y. Gu *et al.*, "Multimodal Attention Network for Trauma Activity Recognition from Spoken Language and Environmental Sound," *IEEE International Conference on Healthcare Informatics*. pp. 1–6, 2019, doi: 10.1109/ICHI.2019.8904713.

[5] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 4960–4964, doi: 10.1109/ICASSP.2016.7472621.

[6] A. Stolcke, "SRILM – An Extensible Language Modeling Toolkit," *Conference of the International Speech Communication Association*. 2002.

[7] S. Jagannath, A. Sarcevic, N. Kamireddi, and I. Marsic, "Assessing the Feasibility of Speech-Based Activity Recognition in Dynamic Medical Settings," *Human Factors in Computing Systems*. 2019, doi: 10.1145/3290607.3312983.

[8] X. Li *et al.*, "Concurrent Activity Recognition with Multimodal CNN-LSTM Structure," *CoRR*, vol. abs/1702.0, 2017.

[9] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *International Conference on Learning Representations*. 2015.

[10] S. Parlak and I. Marsic, "Detecting Object Motion Using Passive RFID: A Trauma Resuscitation Case Study," *ieee Trans. Instrum. Meas.*, vol. 62, no. 9, pp. 2430–2437, 2013, doi: 10.1109/TIM.2013.2258772.

[11] S. Parlak, S. Ayyer, Y. Y. Liu, and I. Marsic, "Design and Evaluation of RFID Deployments in a Trauma Resuscitation Bay," *ieee J. Biomed. Heal. informatics*, vol. 18, no. 3, pp. 1091–1097, 2014, doi: 10.1109/JBHI.2013.2283506.

[12] X. Li *et al.*, "Activity recognition for medical teamwork based on passive RFID," *International Conference on RFID*. pp. 1–9, 2016, doi:

10.1109/RFID.2016.7488002.

[13] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei, “Large-Scale Video Classification with Convolutional Neural Networks,” *Computer Vision and Pattern Recognition*. pp. 1725–1732, 2014, doi: 10.1109/CVPR.2014.223.

[14] X. Li *et al.*, “Region-based Activity Recognition Using Conditional GAN,” *ACM Multimedia*, vol. 2017. pp. 1059–1067, 2017, doi: 10.1145/3123266.3123365.

[15] A. Vaswani *et al.*, “Attention is All you Need,” in *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Eds. Curran Associates, Inc., 2017, pp. 5998–6008.

[16] B. Boashash, *Time-frequency signal analysis and processing: a comprehensive reference*. Academic Press, 2015.

[17] W. Han, C.-F. Chan, C.-S. Choy, and K.-P. Pun, “An efficient MFCC extraction method in speech recognition,” *International Symposium on Circuits and Systems*. 2006, doi: 10.1109/ISCAS.2006.1692543.

[18] M. F. McKinney and J. Breebaart, “Features for audio and music classification,” *International Symposium/Conference on Music Information Retrieval*. 2003.

[19] O. Abdel-Hamid, A.-R. Mohamed, H. Jiang, L. Deng, G. Penn, and D. Yu, “Convolutional neural networks for speech recognition,” *ieee Trans. audio speech Lang. Process.*, vol. 22, no. 10, pp. 1533–1545, 2014, doi: 10.1109/TASLP.2014.2339736.

[20] K. Koutini, H. Eghbal-zadeh, M. Dorfer, and G. Widmer, “The Receptive Field as a Regularizer in Deep Convolutional Neural Networks for Acoustic Scene Classification,” *European Signal Processing Conference*. pp. 1–5, 2019, doi: 10.23919/EUSIPCO.2019.8902732.

[21] K. Kumar, C. Kim, and R. M. Stern, “Delta-spectral cepstral coefficients for robust speech recognition,” *International Conference on Acoustics, Speech, and Signal Processing*. pp. 4784–4787, 2011, doi: 10.1109/ICASSP.2011.5947425.

[22] S. Hershey *et al.*, “CNN architectures for large-scale audio classification,” *International Conference on Acoustics, Speech, and Signal Processing*. pp. 131–135, 2017, doi: 10.1109/ICASSP.2017.7952132.

[23] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” *Computer Vision and Pattern Recognition*. pp. 770–778, 2016, doi: 10.1109/CVPR.2016.90.

[24] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, “Densely Connected

Convolutional Networks," *Computer Vision and Pattern Recognition*. pp. 2261–2269, 2017, doi: 10.1109/CVPR.2017.243.

[25] H. Eghbal-zadeh, B. Lehner, M. Dorfer, and G. Widmer, "A hybrid approach with multi-channel i-vectors and convolutional neural networks for acoustic scene classification," *European Signal Processing Conference*. pp. 2749–2753, 2017, doi: 10.23919/EUSIPCO.2017.8081711.

[26] S. Ioffe and C. Szegedy, "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," Feb. 2015.

[27] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[28] D. P. Kingma and J. L. Ba, "Adam: A Method for Stochastic Optimization," *International Conference on Learning Representations*. 2015.

[29] M. Abadi *et al.*, "Tensorflow: a system for large-scale machine learning.," in *OSDI*, 2016, vol. 16, pp. 265–283.

[30] B. Zoph, V. Vasudevan, J. Shlens, and Q. V Le, "Learning Transferable Architectures for Scalable Image Recognition," *Computer Vision and Pattern Recognition*. pp. 8697–8710, 2018, doi: 10.1109/cvpr.2018.00907.

[31] X. Li *et al.*, "Online process phase detection using multimodal deep learning," *Ubiquitous Computing*. pp. 1–7, 2016, doi: 10.1109/UEMCON.2016.7777912.

[32] A. Graves, "Generating Sequences With Recurrent Neural Networks," Aug. 2013.

[33] P. C. Loizou, *Speech Enhancement: Theory and Practice*, 2nd ed. Boca Raton, FL, USA: CRC Press, Inc., 2013.

[34] J. Benesty, S. Makino, and J. Chen, "Speech enhancement," Springer, 2005, p. 406.

[35] M. Berouti, R. Schwartz, and J. Makhoul, "Enhancement of speech corrupted by acoustic noise," in *ICASSP '79. IEEE International Conference on Acoustics, Speech, and Signal Processing*, 1979, vol. 4, pp. 208–211, doi: 10.1109/ICASSP.1979.1170788.

[36] J. S. Lim and A. V. Oppenheim, "All-Pole Modeling of Degraded Speech," *IEEE Trans. Acoust.*, vol. 26, no. 3, pp. 197–210, Jun. 1978, doi: 10.1130/GES00795.1.

[37] Y. Ephraim, "Statistical-model-based speech enhancement systems," *Proc. IEEE*, vol. 80, no. 10, pp. 1526–1555, 1992, doi: 10.1109/5.168664.

[38] M. Dendrinos, S. Bakamidis, and G. Carayannis, "Speech enhancement from

noise: A regenerative approach," *Speech Commun.*, vol. 10, no. 1, pp. 45–57, Feb. 1991, doi: 10.1016/0167-6393(91)90027-Q.

[39] Y. Ephraim and D. Malah, "Speech Enhancement Using a Minimum Mean-Square Error Short-Time Spectral Amplitude Estimator," *IEEE Trans. Acoust.*, vol. 32, no. 6, pp. 1109–1121, Dec. 1984, doi: 10.1109/TASSP.1984.1164453.

[40] I. Cohen and B. Berdugo, "Speech enhancement for non-stationary noise environments," *Signal Processing*, vol. 81, no. 11, pp. 2403–2418, Nov. 2001, doi: 10.1016/S0165-1684(01)00128-1.

[41] K. W. Wilson, B. Raj, P. Smaragdis, and A. Divakaran, "Speech denoising using nonnegative matrix factorization with priors," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2008, pp. 4029–4032, doi: 10.1109/ICASSP.2008.4518538.

[42] J. L. Durrieu, A. Ozerov, C. Févotte, G. Richard, and B. David, "Main instrument separation from stereophonic audio signals using a source/filter model," in *European Signal Processing Conference*, 2009, pp. 15–19, doi: 10.1099/13500872-142-12-3337.

[43] Y. Xu, J. Du, L.-R. Dai, and C.-H. Lee, "A regression approach to speech enhancement based on deep neural networks," *IEEE/ACM Trans. Audio, Speech Lang. Process.*, vol. 23, no. 1, pp. 7–19, 2015.

[44] Z. Zhang, F. Ringeval, J. Han, J. Deng, E. Marchi, and B. Schuller, "Facing Realism in Spontaneous Emotion Recognition from Speech: Feature Enhancement by Autoencoder with LSTM Neural Networks," *Interspeech 2016*, pp. 3593–3597, 2016.

[45] H. Erdogan, J. R. Hershey, S. Watanabe, and J. Le Roux, "Deep Recurrent Networks for Separation and Recognition of Single-Channel Speech in Nonstationary Background Audio," in *New Era for Robust Speech Recognition*, Cham: Springer International Publishing, 2017, pp. 165–186.

[46] N. Shah, H. A. Patil, and M. H. Soni, "Time-Frequency Mask-based Speech Enhancement using Convolutional Generative Adversarial Network," in *Proceedings, APSIPA Annual Summit and Conference*, 2018, vol. 2018, pp. 12–15.

[47] D. S. Williamson, Y. Wang, and D. Wang, "Complex ratio masking for joint enhancement of magnitude and phase," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 5220–5224, doi: 10.1109/ICASSP.2016.7472673.

[48] M. H. Soni, N. Shah, and H. A. Patil, "Time-Frequency Masking-Based Speech Enhancement Using Generative Adversarial Network," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5039–5043.

[49] K. Paliwal, K. Wójcicki, and B. Shannon, "The importance of phase in speech enhancement," *Speech Commun.*, vol. 53, no. 4, pp. 465–494, Apr. 2011, doi: 10.1016/J.SPECOM.2010.12.003.

[50] N. Takahashi, P. Agrawal, N. Goswami, and Y. Mitsufuji, "PhaseNet: Discretized Phase Modeling with Deep Neural Networks for Audio Source Separation," in *Proc. Interspeech 2018*, 2018, pp. 2713–2717.

[51] S.-W. Fu, Y. Tsao, X. Lu, and H. Kawai, "Raw waveform-based speech enhancement by fully convolutional networks," in *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2017, pp. 006–012, doi: 10.1109/APSIPA.2017.8281993.

[52] S. Pascual, A. Bonafonte, and J. Serra, "SEGAN: Speech enhancement generative adversarial network," in *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, 2017, vol. 2017-Augus, pp. 3642–3646, doi: 10.21437/Interspeech.2017-1428.

[53] A. van den Oord *et al.*, "WaveNet: A Generative Model for Raw Audio," in *9th ISCA Speech Synthesis Workshop*, 2016, p. 125.

[54] K. Qian, Y. Zhang, S. Chang, X. Yang, D. Florncio, and M. Hasegawa-Johnson, "Speech Enhancement Using Bayesian Wavenet," in *Proc. Interspeech 2017*, 2017, pp. 2013–2017, doi: 10.21437/Interspeech.2017-1672.

[55] D. Rethage, J. Pons, and X. Serra, "A Wavenet for Speech Denoising," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 5069–5073, doi: 10.1109/ICASSP.2018.8462417.

[56] F. G. Germain, Q. Chen, and V. Koltun, "Speech Denoising with Deep Feature Losses," in *Proc. Interspeech 2019*, 2019, pp. 2723–2727, doi: 10.21437/Interspeech.2019-1924.

[57] D. Stoller, S. Ewert, and S. Dixon, "Wave-U-Net: A Multi-Scale Neural Network for End-to-End Audio Source Separation.," *Int. Symp. Music Inf. Retr.*, pp. 334–340, 2018.

[58] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," *Med. image Comput. Comput. Assist. Interv.*, pp. 234–241, 2015.

[59] H. Shen, D. George, E. A. Huerta, and Z. Zhao, "Denoising Gravitational Waves with Enhanced Deep Recurrent Denoising Auto-encoders," in *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2019, vol. 2019-May, pp. 3237–3241, doi: 10.1109/ICASSP.2019.8683061.

[60] Z.-H. Ling, Y. Ai, Y. Gu, and L.-R. Dai, "Waveform Modeling and Generation

Using Hierarchical Recurrent Neural Networks for Speech Bandwidth Extension," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 26, pp. 883–894, 2018.

[61] C. Valentini-botinhao, X. Wang, S. Takaki, and J. Yamagishi, "Investigating RNN-based speech enhancement methods for noise-robust Text-to-Speech," *9th ISCA Speech Synth. Work.*, pp. 159–165, 2016, doi: 10.21437/SSW.2016-24.

[62] C. Veaux, J. Yamagishi, and S. King, "The voice bank corpus: Design, collection and data analysis of a large regional accent speech database," in *2013 International Conference Oriental COCOSDA Held Jointly with 2013 Conference on Asian Spoken Language Research and Evaluation, O-COCOSDA/CASLRE 2013*, 2013, pp. 1–4, doi: 10.1109/ICSDA.2013.6709856.

[63] J. Thiemann, N. Ito, and E. Vincent, "The diverse environments multi-channel acoustic noise database: A database of multichannel environmental noise recordings," *J. Acoust. Soc. Am.*, vol. 133, no. 5, pp. 3591–3591, May 2013, doi: 10.1121/1.4806631.

[64] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[65] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," *Proc. Int. Conf. Artif. Intell. Stat. (AISTATS'10). Soc. Artif. Intell. Stat.*, 2010.

[66] A. M. Saxe, J. L. McClelland, and S. Ganguli, "Exact solutions to the nonlinear dynamics of learning in deep linear neural networks," *International Conference on Learning Representations*. 2014.

[67] T. Tieleman and G. Hinton, "Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude," *COURSERA Neural networks Mach. Learn.*, vol. 4, no. 2, pp. 26–31, 2012.

[68] F. Chollet and others, "Keras." 2015.

[69] C. H. Taal, R. C. Hendriks, R. Heusdens, and J. Jensen, "An Algorithm for Intelligibility Prediction of Time–Frequency Weighted Noisy Speech," *IEEE Trans. Audio. Speech. Lang. Processing*, vol. 19, no. 7, pp. 2125–2136, 2011, doi: 10.1109/TASL.2011.2114881.