Published in final edited form as:

IEEE Int Conf Healthc Inform. 2017 August ; 2017: 239–247. doi:10.1109/ICHI.2017.50.

Language-Based Process Phase Detection in the Trauma Resuscitation

Yue Gu[1], **Xinyu Li**[1], **Shuhong Chen**[1], **Hunagcan Li**[1], **Richard A. Farneth**[2], **Ivan Marsic**[1], and **Randall S. Burd**[2]

[1]Department of Electrical and Computer Engineering, Rutgers University, Piscataway, NJ, USA

[2]Trauma and Burn Surgery, Children's National Medical Center, Washington, DC, USA

Abstract

Process phase detection has been widely used in surgical process modeling (SPM) to track process progression. These studies mostly used video and embedded sensor data, but spoken language also provides rich semantic information directly related to process progression. We present a long-short term memory (LSTM) deep learning model to predict trauma resuscitation phases using verbal communication logs. We first use an LSTM to extract the sentence meaning representations, and then sequentially feed them into another LSTM to extract the meaning of a sentence group within a time window. This information is ultimately used for phase prediction. We used 24 manually-transcribed trauma resuscitation cases to train, and the remaining 6 cases to test our model. We achieved 79.12% accuracy, and showed performance advantages over existing visual-audio systems for critical phases of the process. In addition to language information, we evaluated a multimodal phase prediction structure that also uses audio input. We finally identified the challenges of substituting manual transcription with automatic speech recognition in trauma resuscitation.

Keywords

process phase detection; verbal communication logs; deep learning; LSTM; semantic representation

I. INTRODUCTION

Surgical Process Modeling (SPM) focuses on modeling surgical workflows, where a process phase is defined as a sequence of executable activities [1]. This kind of modeling is used for medical education, evaluating team performance, operation planning, and task detection. However, to create high-performance decision support systems, it is necessary to have contextual awareness of the performed activities in relation to the entire process [2][3][4]. To detect and predict process phase, previous research has proposed methods using different data sources. To avoid the labor-intensive manual logging of human observations, sensors (including cameras, wearable sensors, and machine recordings) have been used for data collection. Tool usage recordings [5], medical equipment signals [6], medical event logs [4] [7], body-worn sensors [3], and multimodal Kinect sensors [1] have previously provided input data for surgical phase detection. However, human language has been overlooked as an

input data source for process modeling. To our knowledge, no other research used language information to support process phase and workflow detection. Verbal communication is particularly informative in the trauma resuscitation scenario, where speech contains rich communication information necessary for trauma team dynamics, cooperation, and leadership [8]. In addition, verbal communication provides direct information regarding the performed task, as opposed to tool usage and body movement that are difficult to detect. For example, when asked to identify the current phase of the resuscitation process, medical experts mainly rely on speech, particularly for activities that do not require use of medical tools, such as Airway Assessment and Neurological Assessment, Visual Inspection-N (nose), and Visual Inspection-M (mouth).

We designed a long-short term memory (LSTM) deep learning model using verbal communication to detect the phase of the trauma resuscitation process. Two shotgun microphones were installed in a trauma room to collect speech data, and the audio was manually transcribed to verbal communication logs. Trauma experts identified five trauma resuscitation phases (pre-arrival, patient arrival, primary survey, secondary survey, and post-secondary survey). To predict these phases, we first embedded the transcribed sentences into word vectors, and used an LSTM to extract sentence representations. Then, each sentence representation was sequentially fed into another LSTM to learn inter-sentence contextual information and predict the trauma resuscitation phase. We recorded 30 trauma resuscitation cases (around 17.5 hours) from the Children's National Medical Center in the US. This study was approved by the Hospital's institutional review board. We used 24 cases to train and the 6 remaining cases to test our model. Our results showed an average 79% phase detection accuracy. We also evaluated the potential extensions of our model with multimodal inputs and automatic speech recognition. Our contributions are:

- A multi-stage LSTM resuscitation phase prediction model that automatically extracts sentence-level representations and inter-sentence contextual information from trauma team communications.
- A case study of multimodal (language and speech) phase prediction and automatic speech recognition during trauma resuscitations, which uncovered the issues that need to be addressed for successful application.
- A dataset with audio and transcripts of actual trauma resuscitations available for future research.

The paper is structured as follows: Section II introduces related work on process phase detection and language research in the medical domain. The data collection and ground truth coding are described in Section III. We describe the model structure in Section IV and its implementation in Section V. The experimental results are presented in Section VI. We provide a discussion of model limitations and future work in section VII. We finally conclude in Section VIII.

II. REALTED WORK

With the fast growth of speech technology and with improvements to multimedia systems, language modeling has become more popular in clinical applications. Most speech and

language related medical research focuses on electronic medical records, parsing and mapping patient information to a coded medical ontology or patient records [9][10][11]. However, there has been little research using speech to detect, evaluate, improve, and support the surgical performance.

Several different strategies have been applied to surgical process modeling. A surgical phase detector predicted the laparoscopic cholecystectomy phase by using machine signals representing tool usage [5][6]. However, this method is hard to apply in settings such as the trauma room, because most of the equipment does not generate digital signals. A passive privacy-preserving RFID based approach was used to detect the progress of a surgical process [3]. However, these sensors could interfere with the work and impact patient safety. Manually recorded activity logs were used to predict the surgical phase by using a decision tree structure [4]. However, manual activity logging is subjective, requires expert knowledge, and is labor-intensive. In a recent study, a multimodal deep learning structure was introduced for predicting the trauma resuscitation phase by using depth video and audio input [2]. The results showed adequate performance in phase detection, but the system struggled with phases in which teamwork video and ambient sounds appeared similar. Our preliminary analysis showed that, although gestures and tool usage appeared similar, verbal communication was distinguishable across different phases. For example, the Neurological Assessment in Primary Survey and Visual Inspection-M (mouth) in Secondary Survey have almost indistinguishable tool usage and visual cues, but very different speech.

To perform phase detection based on language information, sentence understanding is necessary. Recent deep learning based techniques showed great performance in sentence semantic analysis. A ConvNet structure was first used as a feature extractor to capture local dependencies between words [12]. To extend to long-range temporal dependencies between words, recurrent neural networks (RNN) and long-short term memory (LSTM) structures were used [13][14]. In addition to learning word relationships, semantic understanding on the paragraph level was also introduced to document analysis. To represent document meaning, CNN-LSTM models were designed for document classification and sentiment analysis [15][16].

For the purpose of this project, we installed a shotgun microphone system in an actual trauma room to capture speech data and designed a LSTM based deep learning model to predict process phase using verbal communication logs. Our model learned both the local and long-range dependencies from language within time windows to predict the phase. We also evaluated a multimodal combination of audio and text, and explored the issues with automatic speech recognition in trauma resuscitations.

III. DATASET

A. Data Collection

The dataset was collected during 30 resuscitations in a trauma room. For safety and convenience, instead of using wearable microphones, we installed two fixed NTG2 Battery or Phantom Powered Condenser shotgun microphones pointed at the team leader group and patient bed area, where most verbal communication takes place (Fig. 1). We recorded the

speech to two channels with 16000Hz sampling rate. Finally, we manually transcribed the audio data for each case.

B. Ground Truth Coding and Data Preprocessing

Five consecutive phases of the resuscitation process were defined by medical experts: pre-arrival, patient arrival, primary, secondary, and post-secondary (Table I). The patient departure phase used in earlier work [2] was omitted because the trauma team usually does not discuss their activities during departure, and because they leave the area covered by microphones. For each sentence, medical experts manually coded the corresponding phase based on the audio-visual records.

We assigned time labels to each sentence corresponding to the time when the last word in the sentence was completed. Even though trauma speech is shorter than regular daily speech and may contain pauses, we still use the last word as the time label to avoid having sentences split over adjacent windows. Next, the sentences were assigned to time windows of 20s audio length. There are two reasons for choosing this window size. First, verbal communication is not uniformly distributed across the duration of the resuscitation. Trauma team members usually speak a lot at the beginning of the primary survey phase, but relatively sparsely during post-secondary phase. Therefore, there might be 30 sentences in one minute during the primary survey phase, but less than 5 sentences in postsecondary survey phase over the same time duration. This fact can also serve as a feature for the phase detection. Second, in the future, our model will be combined with the visual and audio data. Most multimodal architectures use time windows to process the visual and acoustic data, and it would be impossible to assign time window sizes based on a fixed number of sentences

The 20-second time window is shifted through the audio second by second. The sentences for which the time labels belong to a given 20-second window will be fed into the system to predict the process phase. For example, the sentences with labels from 10s to 30s will be fed into the model sequentially to predict the phase at 30s. After this, the same procedure will be repeated for predicting the process phase between 11s and 31s. Example sentences with their time labels are shown in Table II. Our dataset contained 7784 sentences in total.

IV. METHOD

A. Overview

Our model is composed of four modules including the word embedding module, sentence representation module, sentence-sequence representation module, and decision making module (Fig. 2):

- Word Embedding: given a sentence, this layer embeds each word using the word2vec dictionary [17] and then feeds the word vectors into the sentence representation module.
- Sentence Representation: to understand sentence meaning, the word vectors are then processed by the sentence representation module to extract sentence features.

- Sentence-sequence (s-sequence) Representation: this module takes the sentence representations within the last 20s window and outputs a fixed-length feature vector that represents inter-sentence contextual features.
- Decision Making: we use softmax as classifier to predict the process phase based on the sentence-sequence representations.

B. Word Embedding Module

We first mapped the words in each sentence into a corresponding word vector [17]. In natural language processing, word vectors are widely used to represent semantic associations in low-dimensional mappings [17][18]. Each sentence is then represented as a matrix with columns of word vectors and rows of words in the sentence. To initialize the word vectors, we selected word2vec as the embedding dictionary. We compared other word embedding dictionaries, including randomly initialized embedding, Glove word vectors [19], and Collobart word vectors [20], but word2vec had the best performance. Each word was then embedded into a 300-dimensional word vector and unknown words were initialized randomly.

C. Sentence Representation Module

The word embedding layer is followed by the sentence representation module. Convolutional neural networks (CNN) and long-short term memory structures were used to capture the semantic meaning of the sentence from inter-word dependencies [12][21]. Although ConvNets learn spatial features, they do not capture the temporal associations between words in sentences. In our model, we used the LSTM structure because of its ability to handle sequences of various lengths and capture long-range associations [15][21].

LSTM is a special recurrent neural network (RNN) that allows input data with varying length and outputs a fixed-length result [22]. It processes sequence data and overcomes the RNN problem of vanishing and exploding gradients [21]. It has been widely used in natural language processing [15][21][22]. Three different gates (including input, output, and forget) and a memory cell are used to generate the hidden state for each input. Let us define the input sequence as $X = \{x_1, x_2,..., x_t\}$, where x is input data and t is the input time step. The activation output of the input gate is:

$$I_t = \sigma\left(W_I x_t + V_I H_{t-1} + b_I\right) \quad (1)$$

where σ represents the sigmoid function, I_t is an input gate, H_t is the final hidden state, and W, V, and b are the learned parameters. The forget gate decides whether the previous memory should be considered in the current state:

$$F_t = \sigma\left(W_F x_t + V_F H_{t-1} + b_F\right) \quad (2)$$

where F_t is a forget gate. The memory gate combines the old memory state with the forget gate and produces the new memory as follows:

$$M_t = F_t \odot M_{t-1} + I_t \odot tanh\left(W_M x_t + V_M H_{t-1} + b_M\right) \quad (3)$$

The final output and hidden state are defined as:

$$O_t = \sigma\left(W_O x_t + V_O H_{t-1} + b_O\right) \quad (4)$$

$$H_t = O_t \odot \tanh\left(M_t\right) \quad (5)$$

where O_t is an output gate, M_t is a memory cell, and $\odot$ is elementwise multiplication. For each input, the structure considers both the current input x_t, previous memory cell $_{-1}$, and previous hidden state H_{t-1}.

We used one LSTM layer to extract the sentence features from word vectors outputted by the word embedding layer. The final hidden state of this LSTM is used as the sentence representation. In our model, the output of the sentence representation module is a 128-dimensional vector for each sentence. The output vectors are used as the input for the sentence-sequence representation module.

D. Sentence-sequence Representation Module

A single sentence considers only the local dependencies, the information at a particular time during the resuscitation. However, trauma language and activity contain temporal associations across the workflow. These sequence features are lost if we predict phase only based on single sentence representations. During trauma resuscitations, the same activity sentences may appear in multiple phases; even trauma experts need to check previous sentences or videos to determine the current phase during their ground truth coding. Hence, learning contextual and temporal information is necessary for phase detection. LSTMs were designed to obtain long-distance dependencies that can be considered for capturing contextual language information [14]. In our model, we selected 20s as the window size to collect the sentences. We first accumulated the sentence representations based on the timeline and then fed them in sequence into the LSTM structure to learn context information. For each second, the LSTM memory cell slides through the previous 20s sentence representations and outputs the last hidden state as the sequence representation. This sequence vector represents the semantic meaning of the phases in the corresponding 20s based on all sentences appearing in the same time duration. We finally connected a softmax layer at the end of the sequence LSTM structure to predict the phase. Figure 2 shows the sentence-sequence LSTM structure and softmax layer.

V. IMPLEMENTATION AND EXTENSIONS

This section has three sub-parts: model training, multimodal structure, and automatic speech recognition. We first provide the detailed training information of the proposed LSTM based

phase prediction model using language log. We also implemented the model into a multimodal prediction structure combining language information with acoustic (audio) information. Lastly, we test our phase prediction model using text inputs transcribed by a speech recognition engine.

A. Model Training

To train and test the model, we used Keras, a high-level TensorFlow-based neural network library [24]. In the word embedding module, we padded all the sentences to the same length (maximum length: 69) and used word2vec as the embedding dictionary (for 300-dimensional embeddings). For both the sentence and the sentence-sequence representation modules, we empirically set all the hidden states to be 128dimensional and used the Adam optimizer to minimize the loss value, with 0.001 as the initial learning rate and momentum parameters 0.99 and 0.999. To avoid overfitting, dropout was applied during training with probability 0.5, and we trained the model with independent cases. We used 80% of the whole dataset (24 cases) to train and the other 20% (6 cases) to test. The phase was predicted based on sentences within 20-second time windows.

B. Multimodal Implementation

As previously mentioned, different collection methods and data sources have different advantages during phase detection. In the trauma room, RFID performs well at detecting tool usage activities [25]. Visual and audio based multimodal systems perform well at patient arrival and departure detection. To make a system that synthesizes all of these advantages, we can combine different data sources together. However, we do not currently have the video and RFID data corresponding to our shotgun microphone data. To evaluate the feasibility and performance of the multimodal text structure, we adapted our model with the shotgun microphone audio source to build a multimodal architecture. The noise and background sounds proved helpful to phase detection, since the noises and patient voices are distinguishable for the pre-arrival, patient arrival, and patient departure phases [2]. Since our sentence window size is 20s, we used the same window size for the audio branch and applied a ConvNet to extract the audio features, with the same architecture and parameters as previous research [26]. We extracted Mel-frequency spectral coefficients (MFSC) from audio data to form the audio input map, avoiding the locality-compromising discrete cosine transform (DCR) compared with Mel-frequency cepstral coefficients (MFCC) [27]. We reshaped the size of each map to 64×256 before feeding them to the ConvNet. It is worth mentioning that there are 6 audio input channels, since we have two shotgun microphones each with 3 input channels. The detailed structure is shown in Figure 3. We trained the text and audio branches respectively and formed joint feature representations by concatenating the text branch sequence vectors and audio feature fusion outputs (from Fully-Connected Layer2 in audio branch). As suggested [28][29], we trained a three layer neural network with a softmax layer as the phase classifier. We used the same 24 cases as training data and the other 6 as testing data. The results are shown in section V.

C. Automatic Speech Recognition

Speech recognition is a potential replacement for human transcription work. In order to apply speech recognition to the trauma scenario, we fed the audio data from our shotgun

microphones into state-of-the-art speech recognition software to automatically collect the speech transcript, and then inputted the speech recognition result into the LSTM model to predict medical phase. The purposes of this experiment are:

- To build an automatic process phase detection system that uses speech recognition technology to automatically translate the speech signals into text.
- To verify the performance of our proposed model, and to analyze the potential extensions and future work for the system.

Instead of using the surgical simulations to test the system performance [3], we directly use the audio data from actual cases, since the trauma room simulations cannot account for all the real environmental noises, which are key factors for the speech recognition performance. We selected 3 cases from the testing dataset that contained relatively clear and loud speech. Three different state-of-the-art speech recognition engines were used in our experiment: Microsoft Bing Speech (MBS), Google Speech API (GS), and CMU Sphinx (CMUS). The MBS and GS required cloud computation, and we could not modify the speech models. Therefore, we directly fed the audio clips (each clip has one sentence) into the speech engine. CMUS provides the source code to build a custom language and acoustic model [30]. We used the trauma transcripts from training data to build the language model and applied Maximum Likelihood Linear Regression (MLLR) to adapt the acoustic model to the trauma environment [31][32]. We used the same method to feed the speech recognition results into the process phase detection model. The results are shown in section V.

VI. EXPERIMENT RESULTS

A. Evaluation of LSTM-LSTM Model

To evaluate the model performance, we compare four different deep learning based textual models:

- CNN based sentence prediction model: In this model, the phase is predicted by single sentences. The word is embedded by the word2vec dictionary and a ConvNet structure is used as the feature extractor. The model is similar to the CNN-non-static model [12].
- LSTM based sentence prediction model: In this model, we predict phase by single sentences using an LSTM as the sentence feature extractor. The model parameters are the same as our sentence representation module.
- CNN-LSTM based model: Using CNN as the sentence feature extractor and applying an LSTM structure as the sentence-sequence representation module.
- Our proposed LSTM based sequential predication model (LSTM-LSTM).

Table III provides the performances of different models. The result shows 41.7% accuracy for the CNN based sentence prediction model and 44.5% accuracy for the LSTM based sentence prediction model. The CNN-LSTM and LSTMLSTM models show significant performance boosts over the CNN model and LSTM model. Since both the CNN-LSTM model and LSTM-LSTM model consider sequence information and encodes the semantic relationships of sentences, it demonstrates that using the contextual information to predict

the phase is effective. LSTM-LSTM model had 79.1% accuracy, better than the 76.4% of the CNN-LSTM model. We also found that the LSTM model had better performance than CNN in representing sentence meaning.

To further evaluate the proposed LSTM-LSTM model, we provide the corresponding confusion matrix in Table IV. The confusion matrix shows that our model does well at predicting the pre-arrival and secondary survey. Our model achieves around 72% and 81% accuracy, outperforming previous work using visual and audio data (48% and 38%) [2]. As we mentioned before, using visual and audio data has advantages at distinguishing pre-arrival, patient arrival, and patient departure phases, since the head count, patient bed area, machine sounds, and trauma medical team positions are very different at these stages. However, the depth images have limited visibility during the primary and secondary survey. In previous work [2], the depth image comes from one side of the patient bed. During the primary and secondary survey, there are several team members surrounding the bed, occluding the camera's view of the patient area. Since the trauma team members and movements are very similar in depth images during the primary and secondary survey, this occlusion makes activity and phase detection difficult. Audio data also faces the same problem. The machine noise is almost the same during the primary and secondary survey. Compared with the visual and audio data, speech is independent of occlusion, and the examining provider typically reports all primary and secondary survey findings out loud as they are completed. Recognizing this speech should therefore provide more reliable information during these phases. Our results also demonstrate that language information performs well at primary and secondary survey prediction, but performs poorly at patient arrival phase. After further analyzing the transcript, we found that the language information is not very clear during arrival; usually, there is a lot of noise as the patient comes in and multiple people speak at the same time. It is very difficult to capture speech at this stage. Compared with the other phases which have relatively fixed speech content, patient arrival has less language information, and the content is very similar to the pre-arrival and primary survey. For example, the trauma team may discuss the patient weight or patient statue summary during pre-arrival. To fix this problem, we believe we can use relatively obvious visual and audio features (patient entering or exiting).

Our proposed model approaches or exceeds the performance of some existing models for similar scenarios (Table V). Compared with using medical machine signals [5], human language is more general and does not need specific medical equipment. Wearable sensors and RFIDs are inconvenient and may interfere with the operation [3]. Our model only uses language as input; the hands-free microphones do not interfere at all. Instead of wearing the RFID readers, a wearable microphone may capture better data and require less human attention. Unlike in work [30], our two shotgun microphones automatically capture speech and do not require staff to stand and record information in the surgical room. Our confusion matrix results also show the great improvement on *primary phase* and *secondary phase* detection compared with the results of previous work [2].

B. Evaluation of Multimodal Application

The multimodal structure that uses text and speech as input data achieved 82.20% accuracy, which is higher than the models that only use text input. The results demonstrate that our LSTM model can be applied in combination with different data sources. The multimodal structure improves the performance of process phase detection in the trauma room. We believe there are several factors deciding the accuracy. First, audio data indeed indicates some distinguishable features for phase detection, and combining text and audio features enables accuracy improvements. Furthermore, we did not find significant improvement from combining different data sources as in previous work [2], since our microphones were shotguns pointed at two specific areas. As mentioned in section III, shotgun microphones deduct the volume from other areas and enhance the sound from a specific direction, meaning we may miss some audio information from the trauma room as a whole. Most of the data is actually either human voices or patient crying. This reduces the impact of the audio features, because talking always happens throughout the entire resuscitation except patient departure. Finally, the duty shift of the trauma team may influence the accuracy. The trauma members are on different teams, which means the audio features are different. For example, female voices are more common than male voices. Because the main composition is human voice in our audio data, the different person's voices may be extracted as features. However, it will not be useful for phase detection.

C. Speech Recognition Exploration

We provide the results of phase prediction using speech recognition from different speech engines. We show the engine type, word error rate (WER), and phase prediction accuracy in Table VI. 903 sentences with the corresponding audio clips have been applied to the speech engines. The best performance is from the Bing speech API, with 56.37% word error rate. However, it has issues identifying some specific medical terminologies. For example, Bing has difficulty recognizing "Bair Hugger", "CT Scan", and "C-spine". CMU Sphinx4 provides us a method to build our own language model that overcomes the terminology problem. It is obvious that all the WERs were high, meaning all the three engines could not correctly recognize the trauma room speech.

From further analysis, we believe there are three main reasons: position of shotgun microphone, the cocktail party problem, and speaking rate. Even with our shotgun microphones directly pointing to the main areas in the trauma room, some operation noises, patient crying, and speaking direction still heavily influence the audio quality. Vital sounds and operation noises occur throughout the entire resuscitation. Since patients are children, crying noises are very common, covering speech from the trauma team. The direction of the shotgun microphones is fixed, so the speech sounds are weakened when the staff speaks at the side or back of the microphone. The cocktail party problem is a big issue for the trauma scenario [33]. Around 80% of our speech overlaps during the trauma resuscitation and 40% have heavy cocktail party problems. This makes speech hard to recognize, since multiple people speak at the same time with similar volumes. From our results, speech recognition engines have a very hard time with this. Only around 5% of speech that have cocktail party problem were recognized correctly by the engines. Speaking rate is another issue. The speed of trauma speech is much faster than that of daily conversation. The average speaking rate in

the trauma room is 3.983 wps, much faster than 2.667 wps for broadcasts and 2 wps for the general conversation. Considering most speech engines were trained on broadcast and daily conversation audio data, it is reasonable to get low accuracy in our scenario. For speech recognition based phase prediction, the highest accuracy was 41.2% from the CMU Sphinx4. We believe the medical terminologies have a great impact on the accuracy compared with the 39.32% from Bing Speech API.

VII. DISCUSSION AND FUTURE WORK

Our results demonstrate that our language log model performs well in phase detection during the primary and secondary surveys. However, it still has inherent limitations. First, it is hard to detect some phases that do not contain a lot of language information. For example, the patient arrival accuracy is only 22.12% (Table IV), since it has very little language information. Trauma staff seldom use speech to indicate patient arrival. Even for those cases that have patient arrival language information, it is usually just a single sentence. The contextual language information of the patient arrival varies by case and overlaps with the primary survey, making it very hard to detect in our model. The patient departure [2] is also difficult to identify by the language information. Patients are moved outside when the trauma team is ready. The position of the trauma team is changed during the patient departure phase. They do not stand around the patient bed or Mayo stand, making it difficult for the microphone to collect data. To improve the system performance, a multimodal deep learning structure consisting of a video, audio, and text branch should be considered. Patient arrival and departure have obvious visual features, and it has already been demonstrated that visual deep learning approaches work well for these phases [2]. Environmental noise also can be used to improve the accuracy of the patient arrival and departure phases, as there are often distinctive movement sounds.

Furthermore, the fusion of different data sources still poses a challenge. Our results demonstrate that our model can be easily combined with the audio branch and improves the performance of phase detection. Several steps should be considered next. More microphones should be added. Instead of focusing on two specific areas, collecting sound from the entire room provides more information. A multimodal model using all branches together, rather than respectively, should also be considered in our scenario. For example, the CNN-LSTM based deep learning structure is a potential method for our scenario [34]. In next, we must build a dataset that involves text, audio, and video together. We only recorded the text and audio data, but a video branch should be considered for the next step.

Lastly, automatic speech recognition must be improved for the trauma scenario. We used manually transcribed resuscitation verbal communication logs. To make a real-time or online system, automatic speech recognition is necessary. In our current stage, we used two hands-free shotgun microphones to collect the speech data. However, the speech recognition accuracy based on these data is low due to the noisy environments and the cocktail party problem. A more complicated automatic speech recognition system including hands-free and wearable microphones, a speech device selection model, voice activation function, noise reduction function, and custom language model should be considered. Instead of just using two microphones pointed to the patient bed area and team leader group area, more shotgun

microphones will improve the data collection and system performance. Even though trauma experts seldom move during the resuscitation, we found the Jr. Resident, who is the most important person in patient bed area and reports the surgical information during entire resuscitation, may move to the right side of the patient or end of the patient bed to check the patient. This causes language collection problems for our current system that suppresses the sound from side and back. More shotgun microphones from different angles should be considered to completely capture the sound. To improve the audio quality and avoid the cocktail party problem, we could place wearable microphones for important roles that have the most meaningful speech, such as Jr. Resident and Charge Nurse. Unlike single microphone systems, multiple microphone systems need a device selection strategy to choose the clearest sound source from multiple channels and use speech activation to detect silence and divide the sentences. As mentioned in section V, it is necessary to build a custom language and acoustic model to overcome the medical terminology, background noise, and speaking rate problem. This is all potential feature work.

VIII. CONCLUSION

In conclusion, we presented a long-short term memory based deep learning model to predict the surgical phase using resuscitation verbal communication logs. The results show that our model achieves 79% average accuracy, comparable to the existing medical systems. Our model improves the accuracy of primary and secondary survey phases in the trauma resuscitation, which are hard to be detect with other sensor based models and systems. Compared with models that only consider local information, our model combines local and contextual information together to improve phase prediction accuracy. Because we do not cover any specific medical terminology during the modeling, our model is applicable to any medical area. We also demonstrated that our model can be easily extended to a multimodal structure, which uses audio and text together. The speech recognition experiments indicate that a custom automatic speech recognition system with noise reduction, wearable or close distance microphones, and custom language model should be considered for automatic process phase detection.

ACKNOWLEDGMENT

The authors would like to thank the reviewers for their valuable comments and suggestions. This research was supported by the National Institutes of Health under Award Number R01LM011834.

REFERENCES

[1]. Mackenzie L, Ibbotson J, Cao C, and Lomax A, “Hierarchical decomposition of laparoscopic surgery: a human factors approach to investigating the operating room environment,” Minimally Invasive Therapy & Allied Technologies, vol. 10, no. 3, pp. 121–127, 2001. [PubMed: 16754003]

[2]. Li X, Zhang Y, Li M, Chen S, Austin FR, Marsic I, and Burd RS, “Online process phase detection using multimodal deep learning,” 2016 IEEE 7th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON), 2016.

[3]. Bardram JE, Doryab A, Jensen RM, Lange PM, Nielsen KLG, and Petersen ST, “Phase recognition during surgical procedures using embedded and body-worn sensors,” 2011 IEEE International Conference on Pervasive Computing and Communications (PerCom), 2011.

[4]. Forestier G, Riffaud L, and Jannin P, “Automatic phase prediction from low-level surgical activities,” International Journal of Computer Assisted Radiology and Surgery, vol. 10, no. 6, pp. 833–841, 2015. [PubMed: 25900340]

[5]. Blum T, Padoy N, Feußner H, and Navab N, “Modeling and Online Recognition of Surgical Phases Using Hidden Markov Models,” Medical Image Computing and Computer-Assisted Intervention – MICCAI 2008 Lecture Notes in Computer Science, pp. 627–635.

[6]. Padoy N, Blum T, Ahmadi S-A, Feussner H, Berger M-O, and Navab N, “Statistical modeling and recognition of surgical workflow,” Medical Image Analysis, vol. 16, no. 3, pp. 632–641, 2012. [PubMed: 21195015]

[7]. Forestier G, Lalys F, Riffaud L, Collins DL, Meixensberger J, Wassef SN, Neumuth T, Goulet B, and Jannin P, “Multi-site study of surgical practice in neurosurgery based on surgical process models,” Journal of Biomedical Informatics, vol. 46, no. 5, pp. 822–829, 2013. [PubMed: 23810856]

[8]. Bergs EA, Rutten FL, Tadros T, Krijnen P, and Schipper IB, “Communication during trauma resuscitation: do we know what is happening?,” Injury, vol. 36, no. 8, pp. 905–911, 2005. [PubMed: 15998511]

[9]. Singh M and Pal TR, “Voice Recognition Technology Implementation in Surgical Pathology: Advantages and Limitations,” Archives of Pathology & Laboratory Medicine, vol. 135, no. 11, pp. 1476–1481, 2011. [PubMed: 22032576]

[10]. Koivikko MP, Kauppinen T, and Ahovuo J, “Improvement of Report Workflow and Productivity Using Speech Recognition—A Follow-up Study,” Journal of Digital Imaging, vol. 21, no. 4, pp. 378–382, 2008. [PubMed: 18437491]

[11]. Derman YD, Arenovich T, and Strauss J, “Speech recognition software and electronic psychiatric progress notes: physicians’ ratings and preferences,” BMC Medical Informatics and Decision Making, vol. 10, no. 1, 2010.

[12]. Kim Y, “Convolutional Neural Networks for Sentence Classification,” Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2014.

[13]. Mikolov T, Kombrink S, Burget L, Cernocky J, and Khudanpur S, “Extensions of recurrent neural network language model,” 2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2011.

[14]. Pichotta K, and Mooney RJ. “Learning Statistical Scripts with LSTM Recurrent Neural Networks.” AAAI, pp. 2800–2806, 2016.

[15]. Tang D, Qin B, and Liu T, “Document Modeling with Gated Recurrent Neural Network for Sentiment Classification,” Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015.

[16]. Wang J, Yu L-C, Lai KR, and Zhang X, “Dimensional Sentiment Analysis Using a Regional CNN-LSTM Model,” Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2016.

[17]. Mikolov T, Sutskever I, Chen K, Corrado GS, and Dean J, “Distributed representations of words and phrases and their compositionality,” Advances in Neural Information Processing Systems. pp. 3111–3119, 2013.

[18]. Socher R, Perelygin A, Wu JY, Chuang J, Manning CD, Ng AY, Potts C, “Recursive deep models for semantic compositionality over a sentiment treebank,” Proceedings of the conference on empirical methods in natural language processing (EMNLP) vol. 1631, p. 1642, 2013.

[19]. Pennington J, Socher R, and Manning CD, “Glove: Global Vectors for Word Representation,” In EMNLP, vol. 14, pp. 1532–1543, 2014.

[20]. Collobert R, Weston J, Bottou L, Karlen M, Kavukcuoglu K, and Kuksa P, “Natural language processing (almost) from scratch,” Journal of Machine Learning Research 12, no. 8: 2493–2537, 2011.

[21]. Wang S and Jiang J, “Learning Natural Language Inference with LSTM,” Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016.

[22]. Hochreiter S and Schmidhuber J, “Long Short-Term Memory,” Neural Computation, vol. 9, no. 8, pp. 1735–1780, 1997. [PubMed: 9377276]

[23]. Sutskever I, Vinyals O, and Le QV, “Sequence to sequence learning with neural networks,” Advances in neural information processing systems, pp. 3104–3112, 2014.

[24]. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z., Citro C, and Corrado GS, “Tensorflow: Large-scale machine learning on heterogeneous distributed systems,” arXiv preprint arXiv: 1603.04467, 2016

[25]. Li X, Zhang Y, Marsic I, Sarcevic A, and Burd RS, “Deep Learning for RFID-Based Activity Recognition,” In Proceedings of the 14th ACM Conference on Embedded Network Sensor Systems CD-ROM, pp. 164175 ACM, 2016.

[26]. Gu Y, Li X, Chen S, Zhang J, and Marsic I, “Speech Intention Classification with Multimodal Deep Learning,” Advances in Artificial Intelligence Lecture Notes in Computer Science, pp. 260–271, 2017.

[27]. Abdel-Hamid O, Mohamed AR, Jiang H, Deng L, Penn G, and Yu D, “Convolutional Neural Networks for Speech Recognition,” IEEE/ACM Transactions on Audio, Speech, and Language Processing. 22, pp. 1533–1545, 2014.

[28]. Ngiam J, Khosla A, Kim M, Nam J, Lee H, and Ng AY, “Multimodal deep learning,” Proceedings of the 28th international conference on machine learning (ICML-11), pp. 689–696, 2011.

[29]. Srivastava N, and Salakhutdinov RR, “Multimodal learning with deep boltzmann machines,” Advances in neural information processing systems, pp. 2222–2230, 2012.

[30]. Favela J, Tentori M, Castro LA, Gonzalez VM, Moran EB, and Martínez-García AI, “Activity recognition for context-aware hospital applications: issues and opportunities for the deployment of pervasive networks,” Mobile Networks and Applications 12, no. 2–3 pp. 155–171. 2007.

[31]. Walker W, Lamere P, Kwok P, B Raj, Singh R, Gouvea E, Wolf P, and Woelfel J, “Sphinx-4: A flexible open source framework for speech recognition,” 2004.

[32]. Huang X, Acero A, Hon HW, and Foreword By-Reddy R, “Spoken language processing: A guide to theory, algorithm, and system development,” Prentice hall PTR, 2001.

[33]. Bronkhorst AW, “The cocktail party phenomenon: A review of research on speech intelligibility in multiple-talker conditions,” Acta Acustica united with Acustica 86, no. 1, pp. 117–128, 2000.

[34]. Gao H, Mao J, Zhou J, Huang Z, Wang L, and Xu W, “Are you talking to a machine? dataset and methods for multilingual image question,” Advances in Neural Information Processing Systems, pp. 2296–2304, 2015.

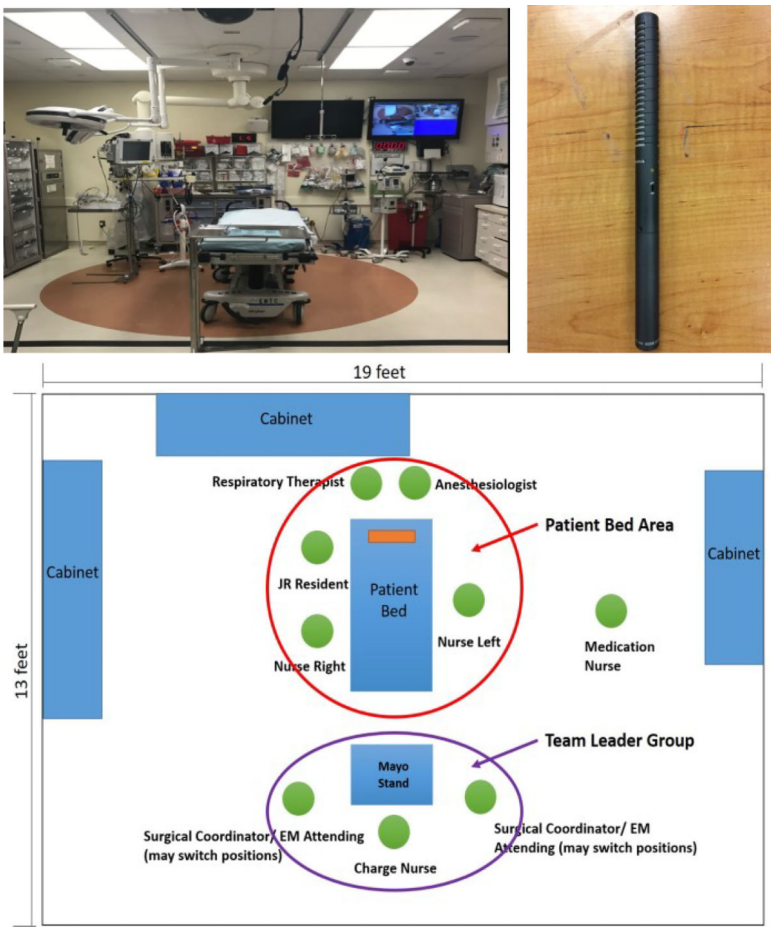

Fig. 1.

Our Hardware Configuration in the Trauma Room. Top Left: Shotgun microphone configuration in the trauma room. Top Right: NTG2 Battery Condenser Shotgun Microphone. Bottom: Trauma room layout and clinical member positions.

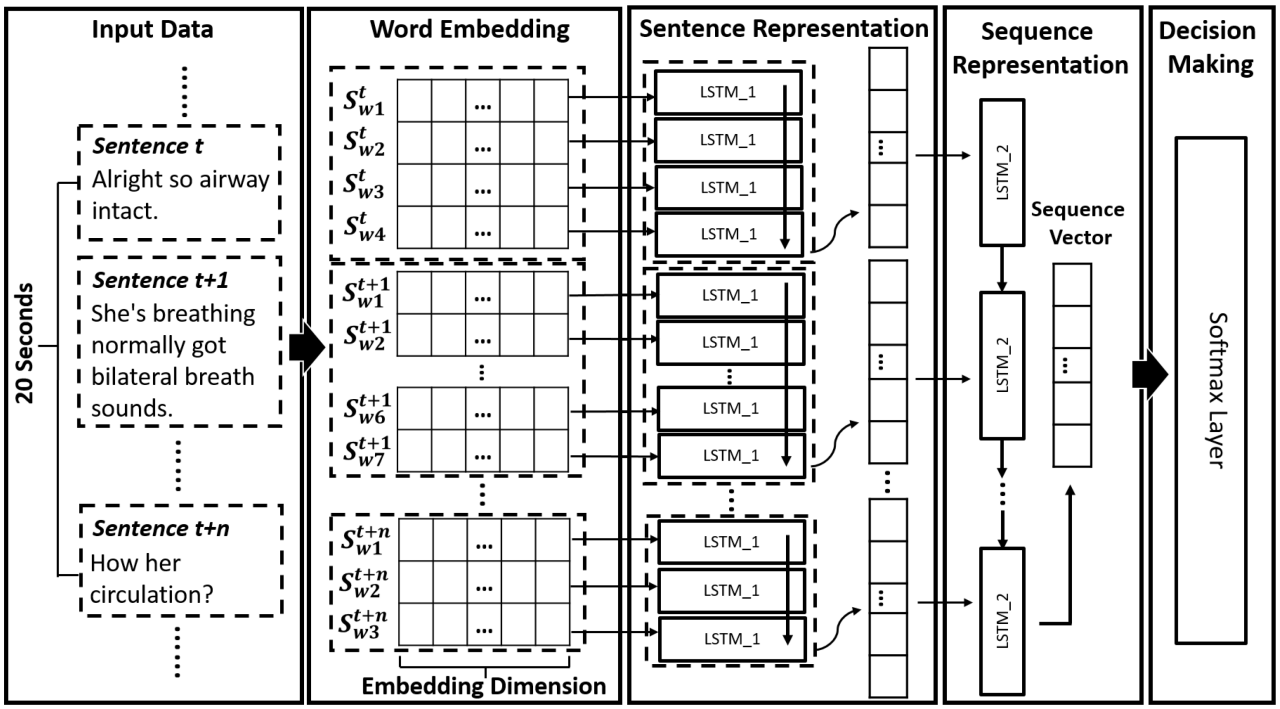

Fig.2.

The proposed LSTM based deep learning model structure. S_w^t: S represents sentence and t is the corresponding number, w represents words in the sentence. Note: we pad all the sentences into the same length.

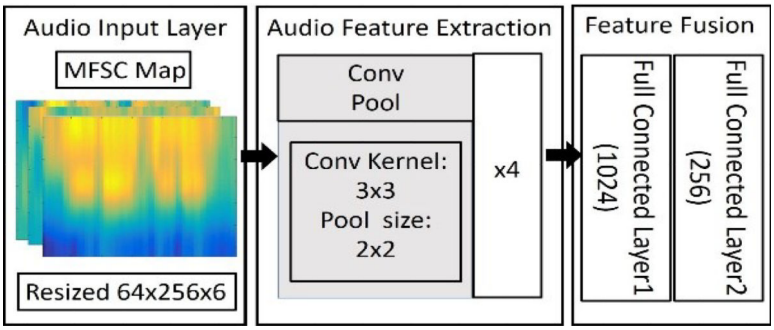

Fig.3.
ConvNet Architecture of Audio Branch

Table I.

TRAUMA RESUSCITATION PHASE DEFINITION

Phase	Definition
Pre-arrival	From the start of the audio recording until patient arrival
Patient Arrival	From patient arrival into the trauma room until the first primary survey activity starts
Primary Survey	From the start of primary survey activity until the first secondary survey activity starts
Secondary Survey	From the start of secondary survey activity until the summary report of the resuscitation
Post Secondary Survey	From the summary report of the resuscitation until the end of the audio data

Table II.

SAMPLE SENTENCE DATA AND TIME LABEL

Phase	Time Label(s)	Sentence
Pre-arrival	115	What's our weight estimate?
Patient Arrival	1193	Alright patient is here.
Primary	1310	Pulse is equal bilaterally.
Secondary	1992	Any stepoffs or deformities you can tell me?
Post Secondary	2278	Is x-ray ready?

Table III.

THE ACCURACY OF FOUR DIFFERENT MODELS

Model	**CNN**	**LSTM**	**CNN-LSTM**	**LSTM-LSTM**
Accuarcy	41.7%	44.5%	76.4%	79.1%

Table IV.

THE CONFUSION MATRIX OF LSTM-LSTM MODEL (PERCENTAGE)

	Prearrival	Patient Arriva	Primary	Secondary	Post-Secondary
Pre-arrival	80.71	14.28	0	0	5.01
Patient Arrival	25.88	22.12	31.76	20.23	0
Primary	9.27	3.12	**72.48**	15.12	0
Secondary	1.26	1.28	8.43	**81.61**	7.41
Post-Secondary	9.15	3.17	0	15.37	72.31

Table V.

THE COMPARISION OF SOME EXISITING MODELS IN SURGICAL SCENARIO

Model	**Data used**	**Accuracy**
Modeling and online recognition of surgical phase using hidden markov models [5]	Medical machine signals	83%
Phase recognition for surgical procedures using embedded and body-worn sensors [3]	Wearable sensors	77%
Activity recognition for contextaware hospital applications [29]	Observation log	75%
Online process phase detection using multimodal deep learning [2]	Video and audio	80%
Our model	**Language log**	**79%**

Table VI.

PERFORMANCE OF PROCESS PHASE DETECTION MODEL BASED ON SPEECH RECOGNITION RESULT

Speech recognition engine	WER	Phase prediction accuracy
Google speech API	63.12%	37.28%
Bing speech API	56.37%	39.32%
CMU Sphinx4	57.35%	41.13%