Public Health and Informatics
J. Mantas et al. (Eds.)

doi:10.3233/SHTI210121

Deep Neural Network Driven Speech Classification for Relevance Detection in Automatic Medical Documentation

Suhail AHAMED [a] , Gabriele WEILER [a, 1], Karl BODEN [b, c], Kai JANUSCHOWSKI [b, c], Matthias STENNES [d], Patrick McCRAE [e], Cornelia BOCK [e], Carina RAWEIN [e], Marco PETRIS [e], Kilian FOTH [e], Kerstin ROHM [a] and Stephan KIEFER [a]

[a] *Fraunhofer Institute for Biomedical Engineering, Sulzbach, Germany*
[b] *Klaus Heimann Eye Research Institute (KHERI), Sulzbach, Germany*
[c] *Eye Clinic Sulzbach, Knappschaftsklinikum Saar, Sulzbach, Germany*
[d] *Fraunhofer Institute for Digital Media Technology, Oldenburg, Germany*
[e] *LangTec, Hamburg, Germany*

Abstract. The automation of medical documentation is a highly desirable process, especially as it could avert significant temporal and monetary expenses in healthcare. With the help of complex modelling and high computational capability, Automatic Speech Recognition (ASR) and deep learning have made several promising attempts to this end. However, a factor that significantly determines the efficiency of these systems is the volume of speech that is processed in each medical examination. In the course of this study, we found that over half of the speech, recorded during follow-up examinations of patients treated with Intra-Vitreal Injections, was not relevant for medical documentation. In this paper, we evaluate the application of Convolutional and Long Short-Term Memory (LSTM) neural networks for the development of a speech classification module aimed at identifying speech relevant for medical report generation. In this regard, various topology parameters are tested and the effect of the model performance on different speaker attributes is analyzed. The results indicate that Convolutional Neural Networks (CNNs) are more successful than LSTM networks, and achieve a validation accuracy of 92.41%. Furthermore, on evaluation of the robustness of the model to gender, accent and unknown speakers, the neural network generalized satisfactorily.

Keywords. Medical documentation, Report generation, Neural Networks, Automatic Speech Recognition, Optical Coherence Tomography.

1. Introduction

The preparation of medical reports, along with other recurrent forms of medical documentation, accounts for a significant portion of a medical practitioner's time [1]. In [2], Smith et al. indicate how documentation is one of the tasks that would be most beneficial for automation in healthcare. Moreover, report generation in a specific

[1] Corresponding author, Fraunhofer IBMT, Joseph-von-Fraunhofer Weg 1, 66280 Sulzbach, Germany; E-mail: gabriele.weiler@ibmt.fraunhofer.de.

clinical setting tends to be repetitive and analogous. For the purpose of this evaluation, the clinical setting studied is for routine medical examinations (including Optical Coherence Tomography (OCT) examinations) of patients undergoing Intra-Vitreal Injections (IVIs).

The implementation of Automatic Speech Recognition (ASR) technology in healthcare, and particularly for medical documentation is quite prevalent and has been scrutinized for over two decades [3]. Even so, such applications are yet to be perfected and produce modest improvements in time and accuracy when compared to the existing norm of dictation and transcription (DT) [4]. A crucial factor that directly affects the turnaround time (TAT) and the number of errors in an automated report is the volume of data that is to be processed by the ASR system. The minimization of the volume of input speech to the information extraction process is one of the key goals of this paper. It is observed that the audio recorded during medical examinations generally consist of a significant volume of speech that is not pertinent to the medical report. The study presented in this paper aims to develop a speech classification module that filters doctor-patient conversations for speech segments that are relevant to the medical report generation process. The proposed speech classification module is intended to be a complementary step implemented to support focused information extraction techniques for automated medical documentation.

Artificial Neural Networks (ANNs) have a long history in speech recognition and other speech-related tasks [5]. The deep neural networks studied in this paper use spectrograms of the speech segments to generate inputs. ANN based speech spectrogram classification models have been implemented in various tasks including Speech Emotion Recognition (SER) and pathology detection from speech with much success [6-7].

2. Methodology

The proposed module is evaluated for classifying speech data recorded during routine medical examinations for patients receiving treatment with IVIs conducted at the Eye Clinic Sulzbach in Sulzbach, Germany. These examinations consist of an anamnesis (present illness, ophthalmological history, general history, medication) followed by a clinical examination (anterior segment, intraocular pressure, posterior segment) and OCT. The dataset is derived from audio recordings of 69 such examinations ranging from 1 to 12 minutes each. The examinations are all conducted in German, and were recorded using high-quality headsets worn by the physician.

The effectiveness of ANNs for a particular application is largely dependent on the architecture of the neurons and a range of parameters, generally referred to as hyperparameters. This evaluation of deep neural networks for the aforementioned use case comprises of comparing the classification performance of different neural network architectures and various topology parameters. Moreover, classification models are trained and validated with characteristic splits in the available dataset to study the dependence of these classifiers on specific speaker attributes.

Review of the literature on spectrogram-based speech classification using ANNs and the experimental comparison of three commonly used frequency domain features, namely Mel Frequency Cepstral Coefficients (MFCCs), log-mel spectrogram features and Constant-Q Transform features, indicated that log-mel spectrogram features are most suitable for this application [8]. A similar study on the Stochastic Gradient

Descent (SGD) and Adam optimizers consistently produced better loss minimization for the latter.

The development of ANNs is programmed in Python using the Keras deep learning library with a Tensorflow backend. Prior to training, the audio data is down-sampled to 20kHz in order to reduce the size of data used for feature extraction while avoiding aliasing of speech. Furthermore, recorded doctor-patient conversations are segmented into utterances, distinguished by pauses in speech, with the help of the speech recognizer engine developed at the Fraunhofer IDMT [9]. The number of training epochs were set to 30, however, the implemented early stopper limited training to between 23 and 26 epochs.

2.1. Dataset

Following the preprocessing of the doctor-patient conversations, the data consisted of a total of 2709 speech segments with a duration of 1 to 14 seconds each. These speech segments are manually annotated, with the label '1' (relevant) for speech segments that contained information that is utilized for the medical report, and '0' (irrelevant) for all other segments. Preliminary exploration of the data indicated that 50.4% of the recorded speech was not relevant for medical documentation.

Furthermore, four characteristic training and validation datasets are generated from the complete data in simple training-prediction data splits. Such data splits were postulated to shed light onto the dependence of the model's performance on speaker attributes. The different datasets maintain the natural class distribution present in the collected data. Table 1 describes the basis of separation and the properties of the training and validation data subsets.

Table 1. Description of the various training and validation data splits. The values in parenthesis indicate the proportion of total speech volume.

Dataset	Criteria for separation	Training data	Validation data
Dataset A	Split by volume; Random	80%	20%
Dataset B	Gender	Female (77%)	Male (23%)
Dataset C	Accent	Native (82%)	Non-native (18%)
Dataset D	Selected speaker	Random speakers (79%)	3 excluded speakers (21%)

2.2. Neural Network Parameters

The architecture of a neural network determines how each neuron processes its inputs, and is a crucial factor in its success. Another important factor that affects the performance of the network is its topology, which includes the number of layers in the network, number of neurons in the network and how they are arranged.

In this paper, the architectures compared are Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs). Both architectures have been quite successful when implemented for applications such as spectrogram classification [9]. Long Short-Term Memory (LSTM) networks and Bi-directional LSTM (BLSTM) networks are the RNNs compared in this study.

3. Results

The results of the neural network optimization process and the dependence of classification performance on the nature of speech are presented in the following sub-sections.

3.1. Neural network performance

The optimization process primarily aims at identifying the most suitable architecture and parameters of the neural network topology. A total of 698 models are trained and validated on Dataset A, through manual, random and grid searches. This includes 340 CNN models, 260 Vanilla LSTM models and 98 Bidirectional LSTM models.

The tested Convolutional Neural Networks consist of 2-10 convolutional layers, 2-10 dense layers and up to 512 neurons per layer. The Recurrent Neural Networks under scrutiny consist of 1-4 LSTM layers, 1-4 dense layers and up to 1024 neurons per layer.

A training-validation data split of 80%-20% is used for training each of the models. Table 2 presents the classification performance and parameter values of the best models of each architecture; the values in parenthesis indicate the number of neurons in each layer.

Table 2. Validation performance and topology of most successful neural networks

Architecture	Functional layers (units)	Dense layers (units)	Validation Accuracy
CNN	3 (50 +75+100)	2 (100 + 100)	0.9241
LSTM	1 (512)	2 (256 + 256)	0.7498
BLSTM	1 (1024)	1 (256)	0.7482

3.2. Speaker dependence

The Convolutional Neural Network, which exhibited highest validation accuracy on the complete dataset (Dataset A), is retrained and validated on the characteristic training and validation data subsets described in Table 1. The validation performance of the CNNs is tracked using four metrics – validation accuracy, precision, recall and Area Under the ROC Curve (AUC). The results for the speaker dependence tests are presented in Table 3.

Table 3. Classification performance of the convolutional network on the different data splits during validation.

Dataset	Accuracy	Precision	Recall	AUC
Dataset B (Gender)	0.9718	0.9472	0.9839	0.996
Dataset C (Accent)	0.9708	0.9545	0.9927	0.9961
Dataset D (Select speakers)	0.9754	0.9658	0.9857	0.987

4. Discussion and Conclusion

In this paper, we have evaluated the implementation of deep neural networks for detecting documentation-relevant speech in a speech corpus recorded during follow-up medical examinations of patients undergoing treatment with IVIs. Although the novelty

of this application and specific setting hinders an exact comparison with other studies, the existing literature advocates cleaning input data for effective ASR [10].

On analysis of the labelled speech data, it is observed that only 49.6% of the speech recorded during these examinations are relevant for medical documentation, hence suggesting a considerable advantage to filtering data prior to implementing complex information extraction techniques. Among the evaluated neural network architectures, CNNs proved to be more successful than LSTM neural networks. Following the optimization of the neural network topology, the most successful CNN model delivered a validation accuracy of 92.41%. This performance supports the potential application of the speech classification module for identifying relevant medical speech.

In the speaker dependence study of the models, the neural network appears to generalize well and exhibits higher accuracy during validation on speech with speaker characteristics excluded in the training data. This unexpected behavior could indicate an insensitivity to the observed speech characteristics. Nonetheless, the obtained results suggest robustness of the classification model to gender, accent and unknown speakers, that is valuable for deploying such a speech classification module which may be centrally trained and utilized by different clinicians in the same organization.

It is noteworthy to mention that this evaluation is limited by its specific clinical setting. Additional research is warranted for such an implementation in a more generic medical setting, for instance, relevance detection in a general physical examination.

Acknowledgement

The ADAPI project is funded by the German Federal Ministry of Economy and Energy (BMWi, Grant id: 16KN071925)

References

[1] Füchtbauer LM, Nørgaard B, Mogensen CB. Emergency department physicians spend only 25% of their working time on direct patient care. Dan Med J. 2013 Jan;60(1):A4558.
[2] Smith K, Smith V, Krugman M, Oman K. Evaluating the impact of computerized clinical documentation. Comput Inform Nurs. 2005 May-Jun;23(3):132-8.
[3] Johnson M, Lapkin S, Long V, Sanchez P, Suominen H, Basilakis J, Dawson L. A systematic review of speech recognition technology in health care. BMC Med Inform Decis Mak. 2014 Oct 28;14:94.
[4] Hodgson T, Coiera E. Risks and benefits of speech recognition for clinical documentation: a systematic review. J Am Med Inform Assoc. 2016 Apr;23(e1):e169-79.
[5] Deng L, Hinton G, Kingsbury B. New types of deep neural network learning for speech recognition and related applications. In: Adams M, Zhao V, editors. IEEE International Conference on Acoustics, Speech and Signal Processing; 2013; Vancouver, Canada. Piscataway: IEEE; 2013. p. 8599-8603.
[6] Shaw A, Vardhan RK, Saxena S. Emotion Recognition and Classification in Speech using Artificial Neural Networks. Int J of Comp App. 2016 July;145(8):5-9.
[7] Rejaibi E, Komaty A, Meriaudeau F, Agrebi S, Othmani A. MFCC-based Recurrent Neural Network for Automatic Clinical Depression Recognition and Assessment from Speech. 2019. arXiv:1909.07208.
[8] Venkataramanan, K, Rajamohan HR. Emotion Recognition from Speech. ArXiv. 2019 Dec;abs/1912.10458.
[9] Huber R, Pusch A, Moritz N, Rennies J, Schepker H, Meyer BT. Objective Assessment of a Speech Enhancement Scheme with an Automatic Speech Recognition-Based System. Speech Communication. In: Doclo S, editor. 13th ITG-Symposium; 2018; Oldenburg, Germany. Berlin: VDE; 2018. p. 1-5
[10] Chiu CC, Tripathi A, Chou K, Co C, Jaitly N, Jaunzeikare D, Kannan A, Nguyen P, Sak H, Sankar A, Tansuwan J, Wan N, Wu Y, Zhang X. Speech recognition for medical conversations. ArXiv e-prints. 2017 Nov;arXiv:1711.07274