THINK INDIA JOURNAL

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

Sentiment Analysis With Fully Supervised Speaker Diarization

[1]Nandini Sethi, [2]Aditi Pandey, [3]Rohit Swami

{[1]nandinisethi2104, [2]aditipandey.ap1,[3]rowhitswami1}@gmail.com

[1,2,3]Department of Computer Science &Engineering, Lovely Professional University, Punjab

ABSTRACT

Speaker diarization with emotions analysis has been one of the most difficult tasks of the Machine Learning and Artificial Intelligence genre. But, this problem is now resolved by some well-known tech companies like Google but depends upon whether they have open sourced their work and not, it made specific population aware of it. In this paper, RNN approach would be discussed in order to provide better accuracy for the sequential data and its advantages over the clustering process, why it best to work with RNN if your data has labels, Using Deep Affects API

We will give the actual visualization of the whole conversation between the people, what were their emotions during their phrases what was their overall sentiment of the conversation involving plenty of emotions in the trained data it gives better accuracy, and even shows it on the visualization page.

Keywords: Sentiment, Diarization, Visualization, RNN

1. INTRODUCTION

Speaker Diarization is a procedure to isolate the sound of various speakers on the basis of the extraction of the features from the audios of speakers. According to google by solving the "who spoke when" we can solve the problem of Speaker Diarization and can assume precisely the real speaker. Speaker Diarization came into the picture in the year of 2006 but showed improvements in huge amount in the year of 2012 and at that time it was not everyone's cup of tea. Speaker Diarization's aim is to detect the timestamps of the different speakers and then to actually classify according to their who it is spoken by. Speaker Diarization has applications in numerous significant situations like in Group discussions analysis that is done manually until now could be totally put on the machine to do, it can also be applied in the Call centres and consultancy departments of the companies which needs to be checked regularly for analysing moreover, it can also be used to analyse the medical conversations. Speaker Diarization is the sensation of recent sensation to make robots and machine closer to human intellect abilities. Speaker Diarization uses a popular model known as Gaussian Mixture Model it assigns frames to the speakers with the help of Hidden Markov Model. There are a lot of online software available in market that can diarize the speakers like ALIZE Speaker Diarization, SpkDiarization, SHoUT developed at the University of Twente to aid speech recognition research, pyAudioAnalysis that provides Feature Extraction, Segmentation, Classification and Applications.

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

WHAT IS RNN?

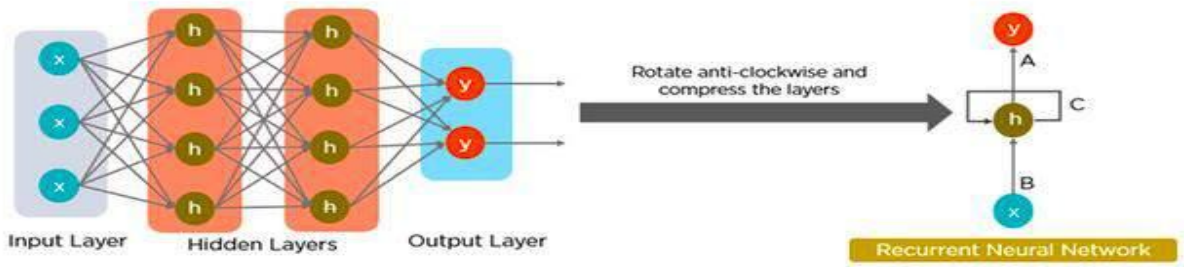

Fig 1: RNN works in the layers which actually loops the information.

Above given diagram, x is given as the input, h is shown as the hidden state values and y is given as the predicted output of the model. In this Figure A, B and C are the parameters of the given network. "In RNN, input of a specific Time stamp(t) is the input of current time t+1 as well as it is the output of the previous hidden state i.e. h(t). So, there is always the looping of information that is given as input at each and every time stamp. Hence, this is called a Recurrent Neural Network (RNN).

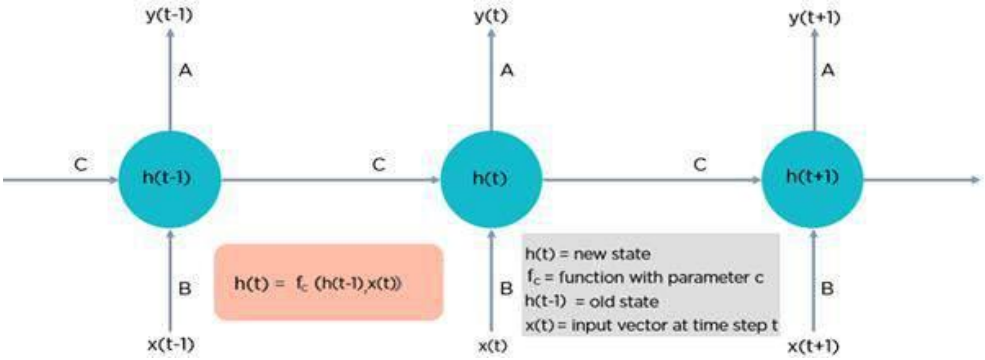

Fig 2: Explanation of RNN system

We can clearly see that at each and every time stamp, the hidden state value of previous is fed as an input to the next state.
RNN can also caption on the real time input streaming of the video and an image based on previous information it has, movements and actions.
Speaker diarization has many applications like: -
1. Medical records: Doctor Vs Patient.
2. Automatic generation of notes in the meetings without any human interventions.
3. Source Speaker Separation.
4. Consulting Firm Data Analysis.

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

Speaker diarization with Sentiment analysis has been a big challenge to the companies,most of the companies either use Speaker diarization or sentimental analysis according to the needs but then every company needs to do either end manually if they are working with the audio data which has been a headache for the companies as it doesn't eliminatethe human effort totally form the process.
Here is one more problem that is not to understand that what diarization is actually. Here are a few points that will make sure to make it understood.
Following are NOT actually Diarization scenarios: -
1. Diarization is not equal to source separation: - No overlap speech.
2. Diarization is equal to speaker change detection: - Labels on change which is different from speaker change detection which is unlabeled.
3. Diarization is not equal to speaker recognition: - Utterances need to be long it is not like small phrases like "Hey Google" or "Ok Google".

Speaker Diarization as it also involves Sentimental Analysis that can also automate the process of analysis of audio files and its application could be in the different fields like Group discussions in which we can analyses the emotions of the participants throughout the discussion and could make the insights about the individual without actually making efforts throughout the process which also ease the process of HR department in any company, similarly it can be used in many places. This project could even be improvised in future to get even the final result without any human intervention which could totally delete the human efforts in the respective fields.

2. PREVIOUS WORK

There is approach that is speaker diarization with LSTM the one we are talking about here is proposed by google, they have worked on the clustering algorithm because it only it is less effective as it can't make good use of supervised data but can be used for the same purpose.
For a long time, I-vector based sound installing methods were the prevailing methodology for speaker check and speaker diarization applications. In any case, reflecting the ascent of profound learning in different areas, neural system-based sound embedding's, otherwise called d-vectors, have reliably exhibited unrivalled speaker confirmation execution. In this approach of Google, they expanded on the accomplishment of d-vector based speaker confirmation frameworks to build up another d-vector based way to deal with speaker diarization. In particular, Google has consolidated LSTM-based d-vector sound embedding's with ongoing work in non-parametric bunching to get a best in class speaker diarization framework. Our framework is assessed on three standard open datasets, proposing that d-vector based diarization frameworks offer huge points of interest over conventional I-vector based frameworks. We accomplished a 12.0% diarization blunder rate on NIST SRE 2000 CALLHOME, while our model is prepared with out-of-space information from voice search logs.

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

2.1 Implementation of the Existing Models

Existing Systems for speaker diarization has been proposed by many leading companies and Machine Learning Scientist but most prominent and remarkable work has been done by Google with the name of "Accurate Online Speaker Diarization with Supervised Learning" there is also a paper written by google on the same by name of "Fully Supervised Speaker Diarization" written by leading scientists John Paisley, Chong Wang, Aonan Zhang, Quan Wang and Zhenyao Zhu.
In whole system developed by Google is implemented by fully supervised model on speaker diarization that gels well with online real-time audio decoding and is very helpful in creating real time subtitles on any platform. Moreover, it could also help in deaf people to largely understand the news on the generic channels which would connect then in mainframe society and largely increase their contribution in it. People of different countries can watch news from different countries which was a difficulty for them due to variation in accent of people. The distinction between Google's model and normal grouping algorithms is that in Google's strategy, every one of the speakers' embedding's is displayed by a "parameter-sharing intermittent neural system (RNN)".

To really see how it functions, consider the model below:-

There are four potential speakers: pink, green blue and yellow,(it is self-assertive - Google's model uses the procedure of a "Chinese restaurant method" to suit the obscure number of speakers). An individual running case of RNN will be created for each section and keeps the RNN state refreshing with new includes from the speaker.

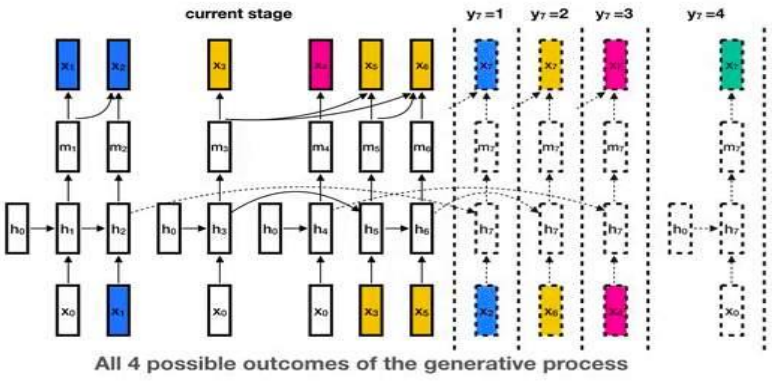

Fig 3: The generative process of the Google model. Colors indicate labels for speaker segments.

In the model below, the blue speaker continues refreshing its RNN state until an alternate speaker

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

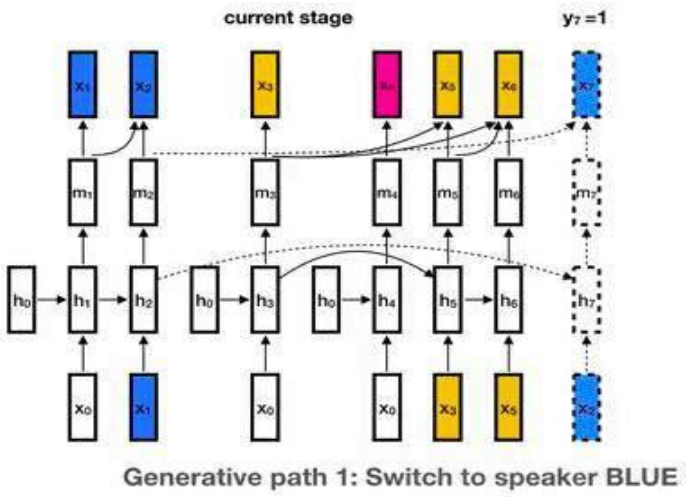

Fig 4: When blue speaks

Same for all others.

3. IMPLEMENTATION OF PROPOSED WORK

Implementation has been done with the help of API for the speaker diarization module other than that Sentimental Analysis are done in both ways that it with SVM and with RNN method but RNN method provides more accuracy so both ways have been implemented here.

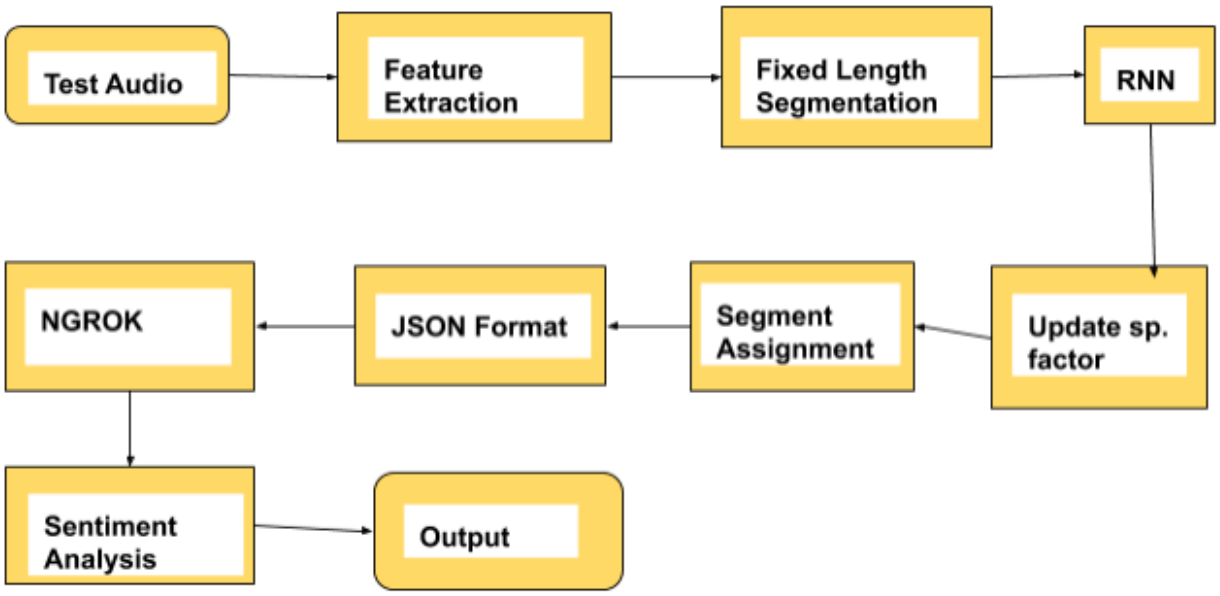

Fig 5: Flowchart of the proposed work.

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

Above flowchart is showing the process in backend that has been followed by the proposed project work and how the Machine Learning Algo works in this model,How API's are working and providing the model greater accuracy and desired results.

Below given is the pseudocode of the proposed work:-
1) Sentiment Analysis with both SVM and RNN.
2) Speaker Diarization with RNN method.
3) Integrate the above modules.
4) Redirect the result to local computer.
5) Visualize the same with .html file with styling of .css file.

There is also a web public API which was created by SeerNet Technologies which has done beautiful and exact same work as Google in this field by they are even providing the services for free and easily accessible for the developers who can't pay for understanding the topics like Speaker diarization and this research paper will talk about the "Sentiment analysis with Fully Supervised Speaker Diarization" through this API integrating it HTML and CSS to make it more understandable to noble audience.
This is how it works and the extra work that has been added to it is the easy visualization for the specific audience to understand.

Figure 5 is a representation of the whole system.

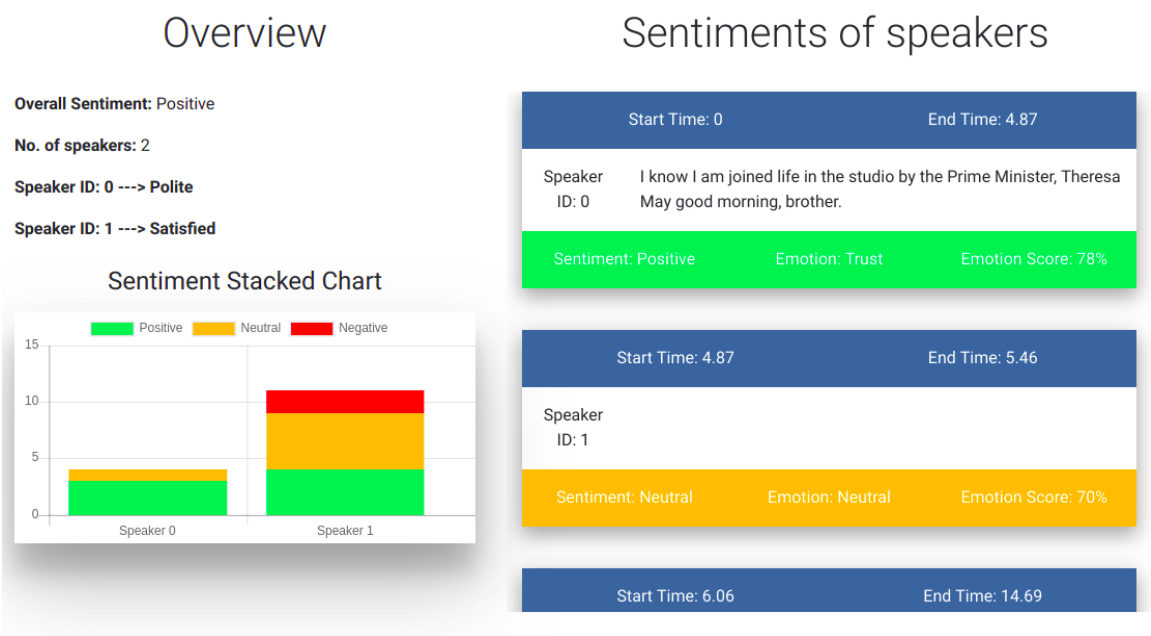

Fig 6: Interface showing various information related to the input audio

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

Above given system has 3 things to analyse that is the number of speakers, emotions of each person with labelled emotions of it and then the stack graph that can be understood easily there is also an overall emotion of the person throughout the conversation that can be used to give the analysis to understand it wholly.

4. RESULTS AND DISCUSSIONS

The proposed model has the following results:-

1) Line by line analysis and the blank space resembles overlapping of two voices or the noisy voices.

Fig 7: Dialogue by Dialogue emotions

Figure 6 shows each dialogue by the labeled speaker that is done by the Algorithm in the model and the output of the model has been shown on the line by line with corresponding color.

This figure shows the accomplishment of the task that was analyzing the sentiments of the speakers with each and every spoken line of them. The space on some places that has been shown in the Software is where the noise of overlapping of the audios has happened.

2) Recognizing the number of speakers, showing the emotions with their corresponding color and even to conclude the overall emotion of that person.

Figure 7 shows the part which shows the valuable information about the input audio.

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

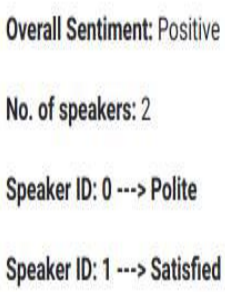

Fig 8: Primary information about the Input File

As the above figure provides us with lots of information like how many participants were there in the input audio stream, Labels of the speakers that is given by the system itself and the overall emotional position of that person throughout the conversation.

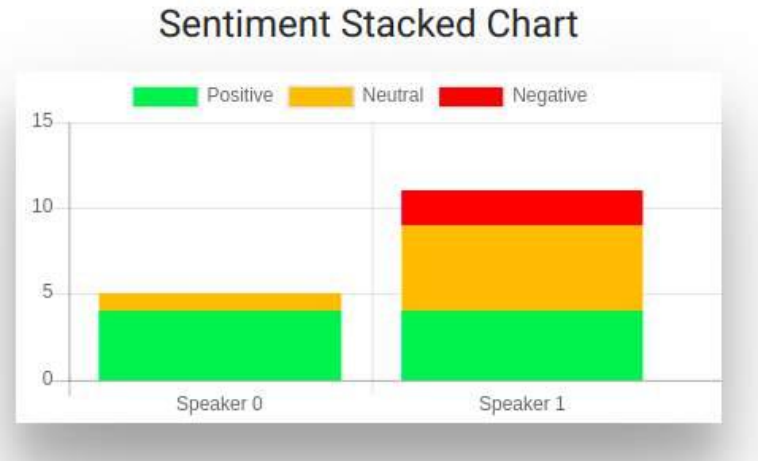

Fig 9: Stack chart showing emotions of labeled speaker

Above given is the figure which is the most understandable and the most useful part of the customer while using this platform. The stack chart actually shows the emotions of the labeled speaker along with the percentage of it in the graph from which would make the user to understand the ratio of each emotion while the conversation.

5. CONCLUSION

Implementation has been done with Deep Affects API which has used RNN in the process rather than the clustering algorithm as it didn't make it suitable of labeled data. RNN proves to be the highly accurate process as it trains by giving feedback to the next cycles. The interface that has been integrated with the API that shows the Number of speakers, the sentences that has been spoken by them, Sentiment related to the sentences, stacked chart related to the emotions of the speaker showing the frequency of corresponding colors of emotions. This interface is easy to understand by even those who are not good with technology; even they can understand and make use of it. It can also be used for the automatic Group Discussion evaluations where there is more

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

than one speaker, System for the Deaf people to make them understand the things going on without the help of anyone. With minor modifications it can also be used in the consulting firms. It can be further modified by providing it data to the concerned problem or a business which can customize the given product. In future this system could be refined and could be trained with many more languages to provide benefit in the local businesses and systems.

REFERENCES

[1] Aonan Zhang, Quan Wang, Zhenyao Zhu1, John Paisley,Chong Wang [2018] FULLY SUPERVISED SPEAKER DIARIZATION
[2] Quan Wang, Carlton Downey, Li Wan, Philip Andrew Mansfield and Ignacio Lopez Moreno [2018] Speaker Diarization with LSTM.
[3] Sue E. Tranter,Douglas A. Reynolds [2006] An Overview of Automatic Speaker Diarization Systems
[4] Xavier AngueraMiro, Simon Bozonnet, Nicholas Evans, Corinne Fredouille, Gerald Friedland, Member, IEEE, and OriolVinyals [2012] Speaker Diarization: A Review of Recent Research
[5] Gregory Sell and Daniel Garcia-Romero, "Speaker diarization with plda i-vector scoring and unsupervised calibration," in Spoken Language Technology Workshop (SLT), 2014 IEEE. IEEE, 2014, pp. 413–417.
[6] Gregory Sell and Daniel Garcia-Romero, "Diarization resegmentation in the factor analysis subspace," in Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on. IEEE, 2015, pp. 4794–4798.
[7] EhsanVariani, Xin Lei, Erik McDermott, Ignacio Lopez Moreno, and Javier Gonzalez-Dominguez, "Deep neural networks for small footprint text-dependent speaker verification," in Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. IEEE, 2014, pp. 4052–4056.
[8] Yu-hsin Chen, Ignacio Lopez-Moreno, Tara N Sainath, MirkóVisontai, Raziel Alvarez, and Carolina Parada, "Locally-connected and convolutional neural networks for small foot-
print speaker recognition," in Sixteenth Annual Conference of the International Speech Communication Association, 2015.
[9] Georg Heigold, Ignacio Moreno, SamyBengio, and Noam Shazeer, "End-to-end text-dependent speaker verification," in Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on. IEEE, 2016, pp. 5115–5119.
[10] F A RezaurRahmanChowdhury, Quan Wang, Li Wan, and Ignacio Lopez Moreno,
"Attention-based models for text-dependent speaker verification," arXiv preprint arXiv:1710.10470, 2017.
[10] Mohammed Senoussaoui, Patrick Kenny, ThemosStafylakis, and Pierre Dumouchel, "A study of the cosine distancebased mean shift for telephone speech diarization," IEEE/ACM

ISSN: 0971-1260
Vol-22-Issue-3-July-2019

Transactions on Audio, Speech and Language Processing (TASLP), vol. 22, no. 1, pp. 217–227, 2014.
[11] Gregory Sell and Daniel Garcia-Romero, "Speaker diarization with plda i-vector scoring and unsupervised calibration," in Spoken Language Technology Workshop (SLT), 2014 IEEE. IEEE, 2014, pp. 413–417.
[12] DimitriosDimitriadis and PetrFousek, "Developing on-line speaker diarization system," pp. 2739–2743, 2017.
[13] Philip Andrew Mansfield, Quan Wang, Carlton Downey, Li Wan, and Ignacio Lopez Moreno, "Links: A high dimensional online clustering method," arXiv preprint arXiv:1801.10123, 2018.
[14] HuazhongNing, Ming Liu, Hao Tang, and Thomas S Huang, "A spectral clustering approach to speaker diarization.," in INTERSPEECH, 2006.
[15] David M Blei and Peter I Frazier, "Distance dependent chinese restaurant processes," Journal of Machine Learning Research, vol. 12, no. Aug, pp. 2461–2488, 2011.
[16] Kyunghyun Cho, Bart Van Merrienboer, CaglarGulcehre, DzmitryBahdanau, FethiBougares, HolgerSchwenk, and YoshuaBengio, "Learning phrase representations using rnn encoder-decoder for statistical machine translation," arXiv preprint arXiv:1406.1078, 2014.
[17] Mark F. Medress, Franklin S Cooper, Jim W. Forgie, CC Green, Dennis H. Klatt, Michael H. O'Malley, Edward P Neuburg, Allen Newell, DR Reddy, B Ritea, et al., "Speech understanding systems: Report of a steering committee," Artificial Intelligence, vol. 9, no. 3, pp. 307–316, 1977.