Predicting the Productivity of a meeting and many other insights from an audio file

Karthik Peddi, Utkarsh Mishra, Nipun Hedaoo

Abstract

This paper proposes an approach that will be able to predict the productivity of a meeting along with many other machine learning insights like speaker positivity comparison, speaker confidence comparison, speaker talk time comparison, Transcript of the conversation, speaker diarization of the conversation, topics that were not discussed in the conversation, topics modelling of the conversation, positive and negative word clouds of the conversation given just an audio file of the recorded conversation. We have also proposed a model where the proposed system can be implemented on a Raspberry Pi platform which takes real time audio recording and does the analysis at the end. The results have been promising when tested with some stock recordings and compared with the ground truth by a communication expert.

Keywords:- Speaker Diarization, Transcribe Audio, Topic modelling, Productivity of meeting, Speaker Confidence, Raspberry Pi, Speech-to-text

Introduction

Every corporate office has a lot of meetings going on through the day, it sometimes can be hectic to maintain a log of whether all the topics that had to be discussed in the meeting had been discussed or not. This is where a system which can predict the productivity of a meeting and list the topics that have not been discussed will shine and be very useful to the employees.

That is why the main idea of this paper is to focus on achieving a product which will be able to do so given an audio file of the recording of the meeting. The paper introduces various machine learning approaches which will be helpful in getting all the machine learning insights from the audio file. This achieved by initially identifying the dialogue of each speaker in the meeting by distinguishing their voices from one another, this is called speaker diarization and this has been achieved in the proposed system using IBM Watson speech to text API.

In the day to day corporate world it is very important to understand your client in a detailed way so as to serve him better next time. To achieve this it is very important to analyse the client's objective discussed during the meeting and have a clear cut idea of what work has to be done and who should it be allocated to. That is why this paper also focuses on these aspects where given the conversation generated from the audio file will be analysed to get relevant details of what work has been given to a particular person and what was the timeline specified etc. Many Machine Learning techniques are used to achieve this in a proper and meaningful manner.

Understanding the client is very essential to how to deal with them, so it is necessary that the positivity of various delegates in the meeting be analysed to know when and where the negativity of the speaker comes into picture and stress on those points to understand the speaker's motive and build a repo with the client.

By analysing the confidence of every speaker that took part in the conversation gives a very good idea of whether the meeting was productive or not. Less confidence in dialogue and many filler words used indicate that the message to be passed out may not have dispersed properly along the delegates, which in turn hurts the entire motive of the meeting, so confidence levels of the dialogues in the conversation is a starting point to furnish the way meetings occur in the company or organization.

All these machine learning insights generated are useless unless the finished product is not able to deliver what it as learnt to everybody in the meeting so as to achieve better results the next time around. So it is very important to send emails to everybody about the various insights gathered in the meeting, which is an important property of a portable system. For this we have proposed a system where a portable device like Raspberry Pi can be used to record live meeting, gather and send all the insights generated by itself to everybody, this has been discussed further later in this paper.

Literature Review

Speaker diarization [1] is an unsupervised task which uses various statistical measures to analyse who spoke when and what they have spoken. Speaker diarization has become a main technology allowing a broad range of functions, including copyright detection, meta-data extraction, data collection, structuring and navigation and document processing. S.E Tranter et al. in [1] have given an overview of the various speaker diarization systems that are being currently used in this domain and which of them are better at recognizing and distinguishing speakers efficiently.

Oku et al. [2] have proposed a speaker diarization system that exploits phonetic information from the audio file to improve better speaker diarization and have a very low latency. In their approach they have initially clustered the speakers based on their acoustic models. They have observed that GMM models also have same number of clusters to that of acoustic models. They used a delta-BIC approach to then segment the speakers. Their system has been tested with data from Japanese Talk shows and it has been proven to be an effective method to use.

Ryo Masumura et al. in [3] have introduced an unsupervised technique which recognizes and distinguishes multiple speakers from the audio file. They also proposed a model which uses a concept they named “role play dialogue topic model” which efficiently utilizes multi party attributes of a conversation. Their proposed topic model realizes a new framework adapts to it efficiently and also considers the language model and trains accordingly.

Emil Șt. Chifu et al. in [4] have presented an approach for aspect based opinion mining which is also an unsupervised machine learning model and uses ant colony optimization. It groups similar sentences into clusters and extracts from every cluster the distinguished aspect of the target object. Their approach is semantically oriented and been tested on a “collection of product reviews”. The performance observed was good and clearly the approach used was dynamic and can be used in any type of topic modelling application.

Sangjun Koo et al. in [5] have proposed a system which tracks the positivity and negativity of various dialogues in a conversation. Their approach uses two methods namely confidence estimation for error modelling and dialogue extraction and abstraction for dialogue state tracking. They have used Phoneme—sequence matching algorithm to estimate confidence for

erroneous Korean input. They were successful at achieving near perfect predictions when deployed at a Korean dialogue prominent centre.

Proposed Work

This section is divided across various sub sections which give a brief description of each module in the paper separately and explain the approach properly. The subsections are Transcribing Audio and Speaker Diarization, Topic Modelling, Positive and Negative Word cloud generation, Positivity comparison of speakers, Productivity estimation of the conversation, Confidence comparison of speakers and Extracting expected outcomes from audio file.

Transcribing Audio and Speaker Diarization

This module deals with the speaker diarization part of the given audio file. We have used a Neural Networks based Machine learning model which learns adaptively and becomes better at recognizing speakers and distinguishing them correctly every time the diarization request is done, this is done by using enhanced models which are accessed through an API created by IBM Watson. Given an audio file the proposed system first transcribes the audio file to generate transcript of the conversation and while doing so takes care of the speaker diarization part so the speakers of each dialogue are distinguished properly.

All the speakers are given tags based on their dialogue and this tag is used to distinguish the dialogue flow of the conversation. The transcription of the audio file is done efficiently and accurately just via an http request and response to the API. The model accurately used its past learnt knowledge to transcribe as accurately as possible.

Topic Modelling

Now that the Transcript of the audio file has been generated it is important to do the topic modelling of the transcript and get all the topics that have been discussed in the meeting. Topic modelling has been achieved by using genism package in python which gets the topics discussed and the LDA visualization of the topics found out is done and saved in the form of an HTML file.

By doing topic modelling one can clearly understand the various topics that have been raised in the conversation which have been identified by various keywords that have been brought up in the conversation. And a LDA visualization of the topics ensures a clear keyword frequency distribution of every single topic in the conversation.

Positive and Negative Word Cloud

Proposed system also generates the positive and negative world cloud from transcript that has been generated from speaker diarization through the Watson API. Positive word cloud of the conversation contains all the positive words that have been uttered in the conversation and classification of a word to be positive/negative is done by checking the frequency of the word in positive dialogues and negative dialogues. The positivity and negativity of the dialogues is extracted by sentiment analysis over each dialogue using Vader sentiment library in python.

Word cloud is a picture containing words in the form of a cloud and the size of the word in the cloud determines the frequency of the occurrence of the words in the conversation. Higher the frequency bigger the text.

Positivity comparison of speakers

The dialogues that have been recognized from the transcript are taken and sentiment analysis is applied on each dialogue to understand whether the dialogue is positive or negative, the dialogue is rated on its positivity and negativity from a scale of 0 to 1 and the positivity of various speakers for their dialogues is compared in a line plot denoting the positivity of each dialogue they uttered.

Sentiment Analysis is the key to this process and is a very good approach used to extract the sentiment of the conversation based on a unsupervised model that extracts the keywords from the conversation and analyses the tone of the conversation based on the keywords combination resulting in a positivity score from the trained machine learning model.

Productivity estimation of the Conversation

The productivity of the conversation is estimated using collection of the topics that have to be discussed in the meeting, prior to the meeting itself and the topics extracted from the transcript are then compared to check which of them have not been raised in the conversation. This is itself is a comprehensive task that requires many modules like topic modelling, similarity detection using synonyms extracted from word nets in python. Then every synonym of the topics to be discussed is matched to topics in the conversation check whether they have been raised in the conversation.

Prior to doing so the stop words from the transcript and the topics to be discussed input must be removed to get accurate productivity. Productivity percentage of the conversation is calculated by the fraction of topics that have been discussed amongst the topics that have to be discussed.

After doing this the topics that have no synonyms present in the conversation are classified as not discussed and prompted to the user, so that it can be addressed at the earliest.

Confidence Comparison of Various speakers

During the speaker diarization part of the proposed system confidence of each dialogue is also extracted from the conversation transcript, this is them used to compute an average of the confidence of each speaker in the meeting and plotted in a way to distinguish their confidence respectively.

Along with the confidence comparison talk time comparison of the speaker is also compared and this is done by checking the amount of time the frequency of the audio file is higher than the typical noise levels of an empty room or a background noise natural to ear.

Extracting action items from an audio file

The proposed system also extracts the expected outcomes that have been mentioned in the conversation. For example, if any of the speakers mention a point of Person A doing this job before this particular time. The following information is extracted from the audio file using

nltk and some subject extraction techniques and this information is stored in a dashboard of a corporate organization in order to evaluate the work.

This module is very useful is such situations and can reduce the workload of making schedules much easier and facile.

Integration with Raspberry Pi

The following proposed system was also implemented on a Raspberry Pi to achieve portability and modularity of the system. The Raspberry Pi module in the system uses a microphone to record real time conversation and perform the analysis and dictate the results to the speaker.

For example, the analysed audio file results in a schedule saying that Person A has to do Job B at Time C. The inbuilt speaker module in Raspberry Pi dictates the above line to the user so that he can get the analysis in a instruction format. All other results of the analysis are also spoken out and suggestions from the analysis are also given to the user for further in achieving their objective.

Implementation Details

Transcription part of the proposed system makes use of the state-of-the-art cloud technology for Speech to text Services i.e. the IBM Watson Speech-To-Text Service.

IBM Watson has the following features:

• Powerful real-time speech recognition

• Highly accurate speech engine

• Built to support various use cases It provides SDKs as well as API references and a pretty good documentation with examples to make using it easier.

Topic modelling is done through the help of gensim, spacy and LDA visualization in python.

Positivity comparison of speakers is done with the help of vaderSentiment analyzer to get the sentiment of each dialogue and plot the positivity of the conversation between the speakers in a line plot using matplotlib.

Confidence comparison of speakers is done at the time of speaker diarization using IBM Watson Speech-to-text API and the confidence of each speaker average is plotted for comparison using matplotlib.

The positive and negative word cloud of the conversation is done using gensim, vaderSentiment and nltk libraries in python.

The action items are extracted using the LSTM and spacy module in python.

The action items are printed in the form of instructions to a specific speaker that it is directed at.

Result and Analysis

We have tested our proposed system by using various excerpts of conversations from various audio file sources, and the proposed system has shown high accuracy in determining the transcript with speaker diarization and predicting the near approximate productivity of the conversation and was also found to be effective in comparing the speakers on various notes like positivity, confidence and talk time. Overall the system is highly rated by the clients and found out to be highly responsive to minor details of the tone of the conversation.

Some of the results that have been gathered from an audio file are shown in figure 1,2,3.

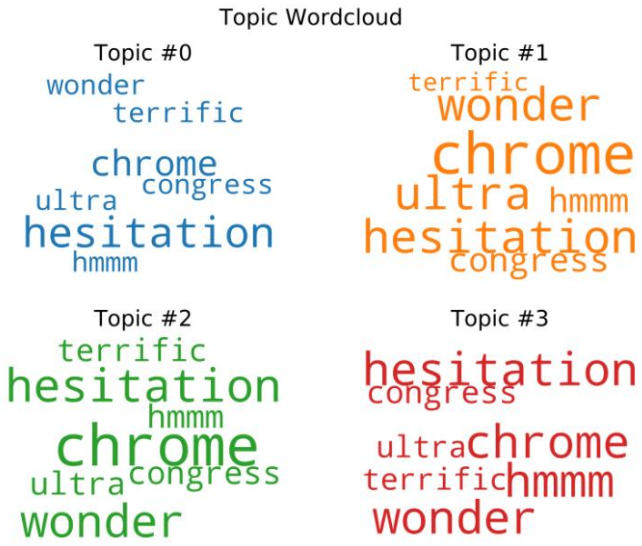

Figure 1: Topic modelling word cloud

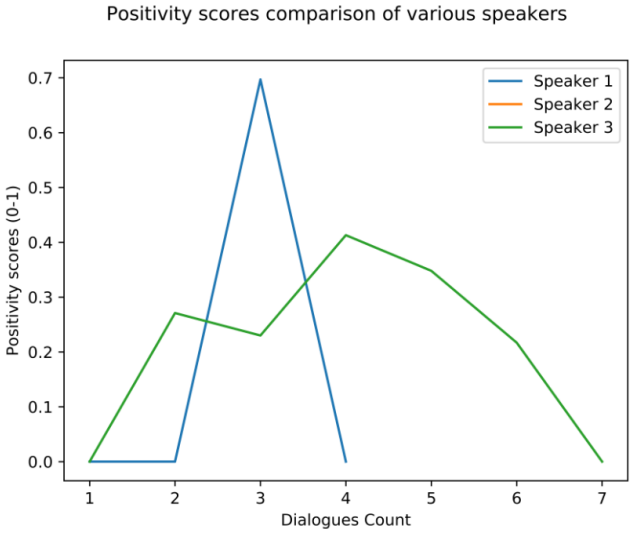

Figure 2: Positivity comparison of various speakers

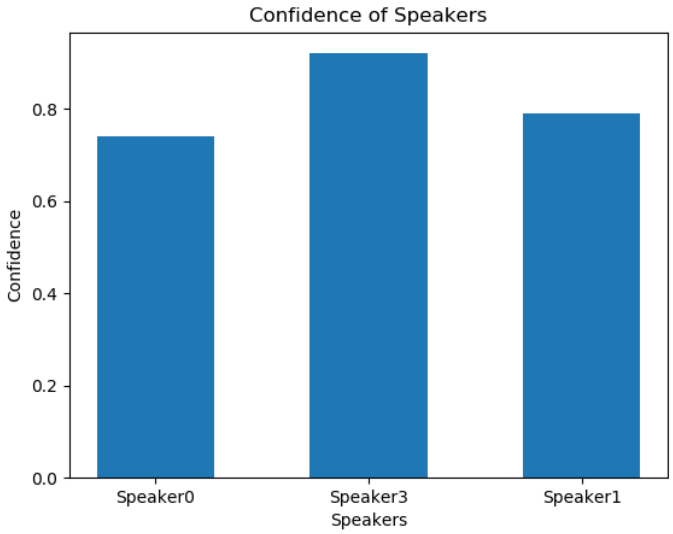

Figure 3: Confidence comparison of various speakers

The following results are an example of how multi speaker audio files can be analysed through the proposed system and brushed through to get the various insights that are very effective in analysing a meeting.

Conclusion

The proposed system has stresses the importance of a system that can analyse the conversation from an audio file and generate insights in a corporate industry where meetings are a daily task. The proposed system is found to be effective in performing various tasks on an audio file that are very helpful in a analysing the productivity of a meeting and do further changes to improve further analysis. This system when deployed in a small scale customer base is found to be effective, accurate and helpful to the work flow of a corporate organization.

Future Work

One of the main topics to focus on in the future in such a domain of meeting analysis, is the continuity of the conversation. It is necessary for corporate delegates to follow a flow of topics in the meeting. For example there might be an initial starting point of the meeting that will be extended with some additional views from various other delegates in the meeting, analysing this properly to check the flow and evaluating it and also accurately predict the productivity is a major challenge and remains unsolved.

References

[1] Tranter, S. E., & Reynolds, D. A. (2006). An overview of automatic speaker diarization systems. IEEE Transactions on audio, speech, and language processing, 14(5), 1557-1565.

[2] Oku, T., Sato, S., Kobayashi, A., Homma, S., & Imai, T. (2012, March). Low-latency speaker diarization based on Bayesian information criterion with multiple phoneme classes. In 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4189-4192). IEEE.

[3] Masumura, R., Oba, T., Masataki, H., Yoshioka, O., & Takahashi, S. (2014, May). Role play dialogue topic model for language model adaptation in multi-party conversation speech

recognition. In 2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4873-4877). IEEE.

[4] Chifu, E. Ş., Leţia, T. Ş., & Chifu, V. R. (2015, October). Unsupervised aspect level sentiment analysis using Ant Clustering and Self-organizing Maps. In 2015 International Conference on Speech Technology and Human-Computer Dialogue (SpeD) (pp. 1-9). IEEE.

[5] Koo, S., Ryu, S., & Lee, G. G. (2015, December). Implementation of generic positive-negative tracker in extensible dialog system. In 2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU) (pp. 798-805). IEEE.

[6] Raaijmakers, S., Truong, K., & Wilson, T. (2008, October). Multimodal subjectivity analysis of multiparty conversation. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 466-474). Association for Computational Linguistics.

[7] https://www.ibm.com/watson/services/speech-to-text/

[8] https://cloud.google.com/speech-to-text/