ISSN: 2229-6956 (ONLINE)
DOI: 10.21917/ijsc.2019.0280

VOICE CALL ANALYTICS USING NATURAL LANGUAGE PROCESSING

V.S. Sudarsan and Govind Kumar
Department of Machine Learning, SeaportAI, India

Abstract

Quality monitoring of calls is a critical activity in call centers. Currently, it is done manually, wherein a person listens to all the recorded audio files or a random sample of audio files to check how the call center representative has performed. Quality monitoring also helps in recording the customer's feedback, which is useful for other business activities like marketing, sales, service, etc. However, this process involves enormous amount of human effort and time besides being error prone. This paper evaluates the application of Machine Learning and Natural Language Processing algorithms in the process of assessing the call center agents.

Keywords:

Voice Call Analytics, Machine Learning, Natural Language Processing, Artificial Intelligence, Artificial Neural Networks

1. INTRODUCTION

In industry 4.0, voicebots are deployed to enhance various workplace processes. Voice bots and chatbots are used in almost all websites. The voice bots are very useful for soft calling in marketing as well as in customer care services. The voicebots are designed to handhold you in reaching your desired result in the most efficient manner. Owing to advancement in AI technology, over the year's voicebots have improved in terms of the quality of interaction and the customer outcomes. Here are a couple of assortments of conversational bots around us. Even our smartphones have chatbots as Siri and Google Assistant. Amazon Alexa is another case of a verbal or voice based chatbot found normally in family units today. These bots fuel the Internet of Things (IoT) ecosystem and help users to naturally engage with all the smart devices at home. It is expected that by the year 2020, 50% of web searches will be voice based. If that is where customers are heading, companies need to be on that platform as well. AI can create a paradigm shift in workplace functioning enabling by voice bots. The old UI-based legacy apps will be replaced by voice bots. There will be no need for different apps for different tasks. The communication and action will become streamlined saving organizational time and other resources. Tech giants like amazon, google, facebook, microsoft are giving more importance to voice-based applications, newly released flagship smart phones have speech to text features, they have released many open source API's, for transcribing speech to text, this shows that the industry 4.0 is moving towards voice based applications and voice-bots. Motivated by this very fact we started evaluation of the voice analytics space. We have used low latency speaker-independent speech recognition, however specific user's voice can be trained. We have chosen the voice call samples from youtube.com which was uploaded for educational purpose. These video files are then converted into audio format. The major challenge in development of this soft-ware is to find the best API among commercially available speech to text API's such as:

- wit.ai
- pocket sphinx
- google

2. RELATED WORK

The basic component of speech recognition is, speech. Speech must be converted from audio to an electrical signal with a microphone, and then to digital data with an analog-to-digital converter. On these digits, several models can be used to transcribe the audio to text. In the above mentioned models we tried google voice, pocket sphinx and wit.ai.

The Most modern speech recognition systems rely on what is known as a Hidden Markov Model (HMM). This methodology chips away at the presumption that a speech signal, when viewed on a shorter scale, can be reasonably approximated as a stationary process that is, a process in which measurable properties do not change with time. The final output of the HMM is a sequence of vectors. To decode the speech into text, gatherings of vectors are coordinated to one or more phonemes (a fundamental unit of speech). This calculation requires training, since the sound of a phoneme varies from speaker to speaker, and even varies from one expression to another by the same speaker. A special algorithm is then applied to determine the most likely word (or words) that produce the given sequence of phonemes. One can imagine that this whole process may be computationally expensive. In many modern speech recognition systems, neural networks are used to simplify the speech signal using techniques for feature transformation and dimensionality reduction before HMM recognition.

We used different techniques for different API's. While using wit.ai we have to convert large audio files to small chunks as wit.ai allows only small audio files for transcribing to avoid high latency

3. SPEECH RECOGNITION

Speech recognition process (Fig.1) deals with speech variability and account for learning the relationship between specific utterance and the corresponding word or word. There are three approaches to speech recognition.

- Pattern Recognition Approach
- Artificial Intelligence Approach
- Acoustic Phonetic Approach

In Acoustic Phonetic approach the speech recognition is based on discovering speech sounds and labelling them.

Pattern recognition involves the training or developing a system which will divide a large number of individual examples into groups called classes. As the audio call is subjected to a lot of noise we use pattern recognition approach according to which the audio is compared with the reference classes and the similarity is measured.

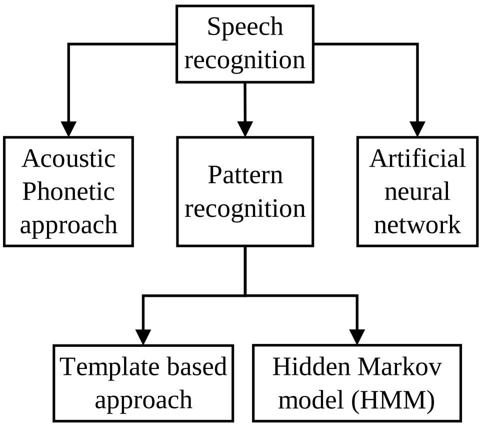
Speech recognition
Acoustic Phonetic approach
Pattern recognition
Artificial neural network
Template based approach
Hidden Markov model (HMM)

Fig.1. Taxonomy of Speech Recognition

In template based approach the patterns are stored as dictionary of words. Discourse is recorded by coordinating an obscure spoken articulation with every one of these reference formats and choosing the class of the best coordinating example. By utilizing this strategy blunders in distinguishing littler words can be eliminated.

Template based approach to speech recognition has provided a family of technique that has advanced the field considerably during the last two decades. This approach is simple. This methodology is straightforward. It is the methodology of coordinating obscure discourse with a lot of pre-recorded words so as to locate the best match. This methodology has the advantage of utilizing splendidly ex-act word models; however, this has the drawback that the pre-recorded formats are foreordained. So difference in speech signals must be demonstrated by utilizing numerous formats per word, which surely ends up unreasonable. The biggest disadvantage of Template training is that it is very costly and it is impossible to train the entire language vocabulary

As discussed earlier, in Hidden Markov Models (HMMs) to decode the speech into text, groups of vectors are matched to one or more phonemes This calculation requires training, since the sound of a phoneme varies from speaker to speaker, and even varies from one utterance to another by the same speaker. A special algorithm is then applied to determine the most likely word (or words) that produce the given sequence of phonemes.

Another approach in acoustic modelling is the use of neural networks. Artificial Neural Networks (ANN) or connectionist systems are computing systems that are inspired by, biological neural networks that constitute animal brains. Such frameworks “learn” to perform errands by thinking about models, by and large without being customized with task-explicit guidelines. Data that floods through the system impacts the structure of the ANN in light of the fact that a neural system alters or learns, it could be said in light of that information and yield.

They are able to solve a lot more problematical recognition tasks, but do not scale as excellent as Hidden Markov Model (HMM) when it comes to huge dictionaries hence they are not used in long audio files. They can deal with low quality, noisy data and speaker independent situations. Such systems can accomplish greater accuracy than HMM based systems, as long as there is training data and the vocabulary is finite. A more common approach using neural networks is phoneme recognition. This is an active field of research, but generally the results are improved than HMMs.

In this paper we have tried three Speech recognition API’s in which sphinx uses acoustic phonetic approach, google uses Hidden Markovian model which is a machine learning algorithm and wit.ai uses artificial intelligence approach.

4. PROCEDURE USED IN WIT.AI

- The audio calls of the particular agent is merged into one large file programmatically in order to optimise the transcribing process
- In order to reduce latency (the delay before a transfer of data begins following an instruction for its transfer) we have to break the file into small chunks (10s).

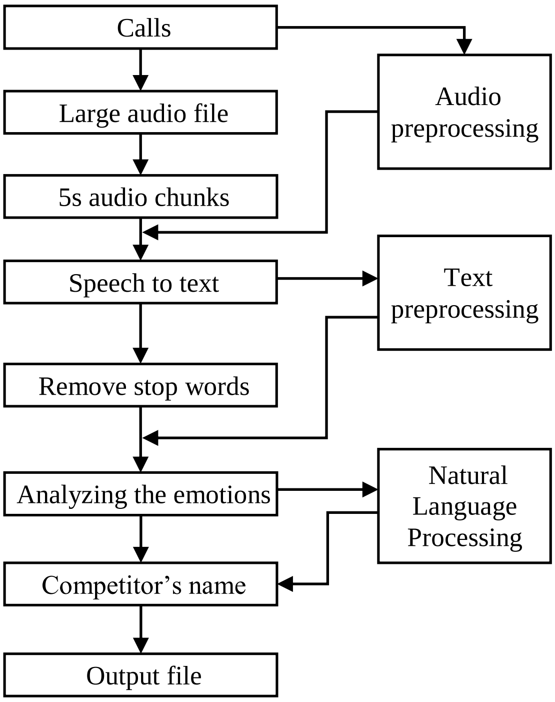

Fig.2. Flow chart for wit.ai

- The chunks are the transcribed.
- The stop words are removed and various emotions are detected using Natural Language Processing.

The Procedure used in sphinx and google:

- The audio calls are first converted to text.
- The stop words are removed and various emotions are detected using Natural Language Processing.

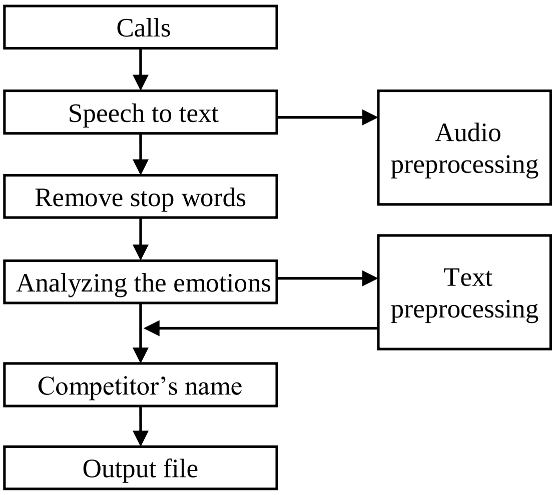

Fig.3. Flow chart for google and sphinx

5. CALL ANALYTICS METRICS

We used fuzzywuzzy module to compare the manually transcribed text file of the voice calls and the text data.

The accuracy of transcribing was tested between pocket sphinx, google cloud speech to text and google for accuracy using Harvard Open Speech repository from the internet. The accuracy of wit.ai was more compared to other packages. This work only focuses on the analysis of call centre conversations and future work will include our own automated speech recognition software. The transcribed text is stored into data sets and the entire code is done using python language.

The pocket sphinx module is used for both offline and online conversion of speech to text and wit.ai is used for online conversion of speech to text hence the accuracy of wit.ai is high.

The following metrics are widely used in the call centre industry, to perform analytics. Transcribed from different API's like sphinx, google cloud, google voice etc.

5.1 CUSTOMER EMOTIONS DETECTION

The various emotions of the customer such as whether he is satisfied or not, whether he is happy or angry with the agent is detected. This data can also be used to gather insights like review for new product launch, marketing, sales, service, etc.

5.2 BANNED WORDS

Use of banned words can be detected. Words in conversation texts have been compared with the list of banned words and checked whether agent or customer has used any banned word or not. The number of matching banned words are recorded as number of banned words in database.

5.3 GREETING WORDS

In industries like call centre, where there is direct interaction with the customer, it is very much important to for the call centre agent to greet the customers. But many times because of the hectic schedule, the representative might miss greeting the customers. Words in conversation texts have been compared with the list of greeting words and checked whether agent or customer has used any greeting word or not. The number of matching greeting words are recorded as number of greeting words in database.

5.4 USAGE OF COMPETITORS NAME

The various cases under which the competitors name is used by the customer can be recorded. This can give deep insights in marketing.

5.5 PERFORMANCE SCORE

A performance score based on the above mentioned parameters is devised to evaluate the call centre representative's performance.

6. EXPERIMENTAL RESULTS

6.1 SPEECH TO TEXT ACCURACY

It can be seen that the speech to text accuracy for wit API is nearly 90% when compared to the other two API's. Google speech to text API has an Accuracy of 87%.

Table.1. Speech to Text Accuracy

API	Accuracy of human emotion 1	Accuracy of human emotion 2	Accuracy of human emotion 1	Accuracy of human emotion 2
Sphinx	38%	63%	Yes	No
Wit	100%	90%	No	Yes
Google	83%	87%	Yes	No

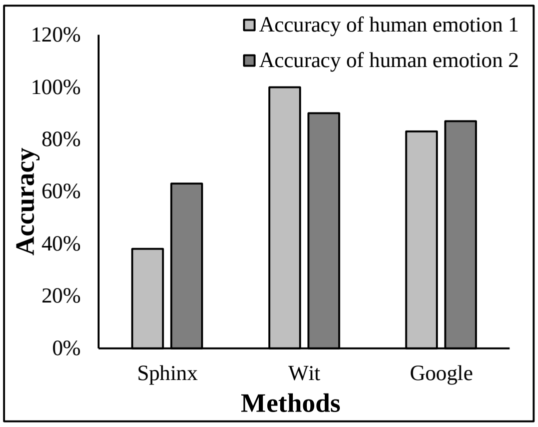

Fig.3. Speech to Text Accuracy

6.2 DETECTION OF EMOTIONS IN A CALL

The percentage of angry calls detected is compared with the actual numbers. The percentage of wit and google is 100%.

Table.2. Angry call percentage

API	Percentage of angry calls
Sphinx	33%
Wit	100%
Google	100%
Actual text	100%

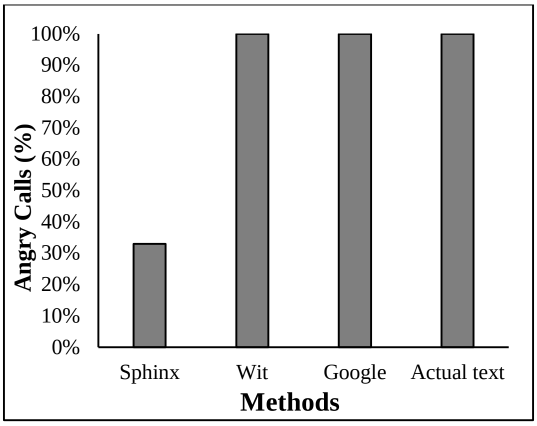

Fig.4. Satisfied call percentage

Table.3. Percentage of happy calls detected

API	Percentage of happy calls
Sphinx	0%
Wit	33%
Google	33%
Actual text	33%

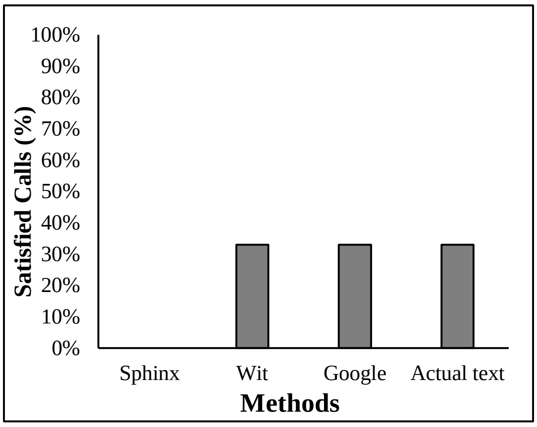

Fig.5. Satisfied call percentage

The actual percentage of happy calls is 33% which is perfectly detected by wit.ai and google, whereas sphinx couldn't identify happy calls.

Table.4. Percentage of unsatisfied calls detected

API	Percentage of calls
Sphinx	33%
Wit	100%
Google	100%
Actual text	100%

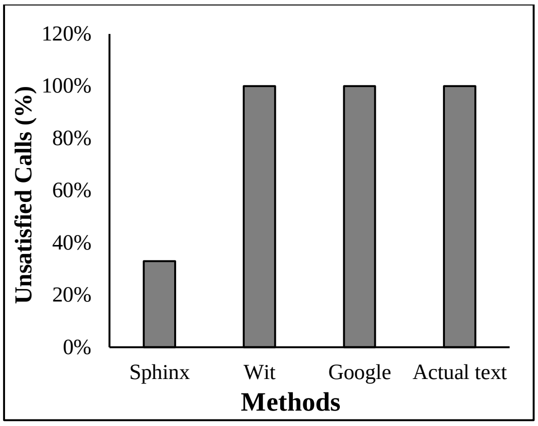

Fig.6. Unsatisfied call percentage

The actual percentage of unsatisfied calls is 33% which is perfectly detected by wit.ai and google, whereas sphinx couldn't identify unsatisfied calls.

Table.5. Percentage of banned words detected

API	Percentage of banned calls
Sphinx	0%
Wit	0%
Google	0%
Actual text	0%

There was no banned words in the calls, so it was not detected.

Table.6. Percentage of competitors name detected

API	Percentage of competitor name in calls
Sphinx	0%
Wit	33%
Google	0%
Actual text	33%

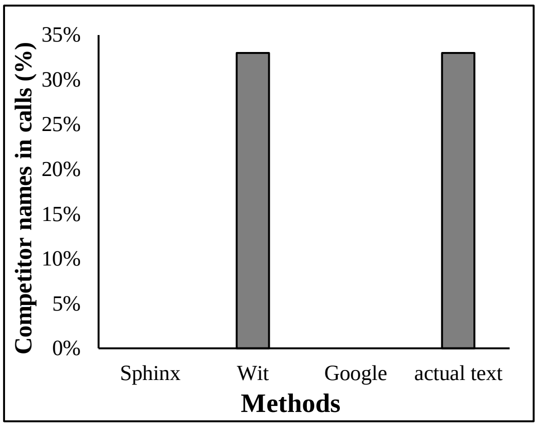

Fig.7. Percentage of competitors name detected

Competitors name being a noun is very hard for API'S to detect in which wit.ai performed a very good job.

7. PERFORMANCE EVALUATION

Table.7. Overall accuracy of Human Emotions Detected

API	Percentage of human emotion in calls
Sphinx	38.33%
Wit	100%
Google	83.33%

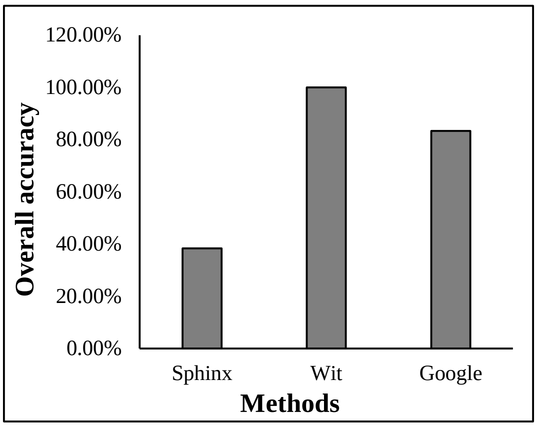

Fig.8. Overall Accuracy of Emotions

From the above mentioned data, we found a cumulative overall accuracy of all human emotions detected in which wit.ai has 100% accuracy google has 83%.

Table.8. Comparison of different API's

API	Accuracy of human emotion 1	Accuracy of human emotion 2	Accuracy of human emotion 1	Accuracy of human emotion 2
Sphinx	38%	63%	Yes	No
Wit	100%	90%	No	Yes
Google	83%	87%	Yes	No

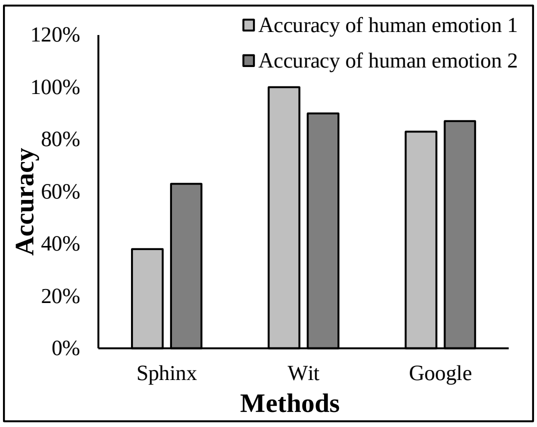

Fig.9. Comparison of different API's

8. CONCLUSIONS AND FUTURE WORK

In this paper we have developed a software to measure performance of call centre representative and also to get useful insights for business analytics. We have compared different API's based on which we can use the API's for different conditions.

We are working on a speech to text convertor using fuzzy mathematics, and deep learning models to increase the accuracy of the existing methods. We are also working on voice pitch based customer emotions.

REFERENCES

[1] S.K. Kopparapu, "*Non-Linguistic Analysis of Call Enter Conversations*", Springer, 2015.

[2] Steven Bird, Ewan Klein and Edward Loper, "*Natural Language Processing with Python*", Oreilly Media, 2009.

[3] Christopher D. Manning and Hinrich Schutze, "*Foundations of Statistical Natural Language Processing*", MIT Press, 1999.

[4] Masamichi Kon, "A Note on Zadeh's Extension Principle", *Applied and Computational Mathematics*, Vol. 4, No. 1-2, pp. 10-14, 2015.

[5] G. Chakraborty, M. Pagolu and S. Garla, "*Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using* SAS", SAS Institute, 2014.

[6] Hironori Takeuchi, L Venkata Subramaniam, Tetsuya Nasukawa and Shourya Roy, "Automatic Identification of Important Segments and Expressions for Mining of Business-Oriented Conversations at Contact Centers", *Proceedings of Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 458-467, 2007.

[7] G. Mishne, D. Carmel, R. Hoory, A. Roytman and A. Soffer, "Automatic Analysis of Call-Center Conversations", *Proceedings of 14th ACM International Conference on Information and Knowledge Management*, pp. 453-459, 2005.

[8] Martine Garnier-Rizet, Gilles Adda, Frederik Cailliau, Sylvie Guillemin-Lanne, Claire Waast-Richard, Lori Lamel, Stephan Vanni and Claire Waast-Richard, "CallSurf: Automatic Transcription, Indexing and Structuration of Call Center Conversational Speech for Knowledge Extraction and Query by Content", *Proceedings of 6th International Conference on Language Resources and Evaluation*, pp. 553-559, 2008.

[9] Ayushi Y. Vadwala, Krina A. Suthar, Yesha A. Karmakar, Nirali Pandya and Bhanubhai Patel, "Survey Paper on Different Speech Recognition Algorithm: Challenges and Techniques", *International Journal of Computer Applications*, Vol. 175, No. 1, pp. 31-38, 2017.