Pipeline for XAI based Automatic Audio Call Audit

Upasana Tiwari*†, Rupayan Chakraborty*†, Sumit Divolia†, Sunil Kumar Kopparapu *†
*TCS Research and Innovation-Mumbai
†Tata Consultancy Services Limited, INDIA

***Abstract*—In this paper, we describe an industry grade, functional, automatic explainable audit system for service desk call conversations. Often these interactions are laced by the various linguistic and non-linguistic cues which become prominent marker in evaluating the overall conversation quality. The implemented system is capable of extracting these cues by careful pipelining and integration of several automatic speech processing engines, namely, Speaker Diarization, Automatic Speech Recognition (ASR), Speech Emotion Recognition (SER), Speaking Rate Monitoring among others. In addition, we incorporate an explainable insight engine into the audit system, and subsequently evaluate 15 high-level conversation quality attributes. The audit system, deployed in cloud, has a web interface allowing (a) auditors to upload bulk calls in a single audit request (b) simultaneous access for multiple auditors (c) time complexity of $0.6\times$ call duration (d) explainable audit attribute evaluations. In summary, the automatic audit system not only enables audit of cent percent of the call conversations but also reduces the time complexity significantly, eliminates human bias, thereby leading to an increase in customer satisfaction and call quality at service desk, in comparison with the manual audit process. On a test bed of 80 real call conversations, we show that the automatic audit is able to achieve 92.25% agreement with the gold standard evaluation.**

***Index Terms*—speech analysis, call quality audit, explainable AI**

I. Introduction

Call center service desk in any enterprise receives thousands of audio call on daily basis. One of the prime focus of any service desk is to keep the customer satisfaction index high. In this pursuit, it is essential to monitor the conversational quality of call-center interactions to keep track of the standard of service offered by agents to their customers. Call audit is one such process that requires evaluation of high-level attributes to measure the quality of a call, either manually or automatically. The major shortcomings with the manual audit process are: (1) time complexity (2) audit expense (3) human-bias (4) lack of consistency in explaining the evaluations.

In literature, there exist systems to automatically monitor the conversation quality of agent-customer interaction so that corrective measures could be taken to keep the customer satisfaction high. The work in [1]–[4] proposed call monitoring system by using text analytics and information retrieval methods. In [5], authors proposed online call monitoring by using finite state machines. In contrast to using text and speech analytics, Pallotta proposed an interaction mining tool built on pragmatic analysis to call-center analytics [6]. In [7], author presented methods to identify problematic calls in call-center, solely based on non-linguistic analysis. Karakus proposed a distributed call quality monitoring system using several qualitative measures [8]. Zweig proposed call quality monitoring system by using speech recognition, pattern matching, and maximum entropy [9]. Though there exist several pursuits in literature to address automatic call quality monitoring, all of them require the need to train (and further tuning) the domain-specific models to perform automation, this is an overhead in terms of availability of data required for the model to generalize well.

To overcome the above mentioned shortcomings, we built an industry grade automatic call audit system, by integrating several state-of-the-art audio processing engines and incorporating an ability to explain the evaluated audit attributes. The audit system caters to both linguistic and non-linguistic information extracted from conversation through a pipeline of several automatic engines: Speaker diarization, ASR, SER, and Speaking Rate Monitoring. Then the integration generates a quality report by evaluating 15 enterprise specific attributes by marking each as Yes/No and provides explanation against each evaluation. The main contributions of this paper are: (i) Incorporation of automatic evaluation of 15 high-level attributes by integrating of 4 speech processing engines (based on AI and Deep Learning) (ii) $6\times$ times faster than manual audit, (iii) 100% call audit in comparison to the very low percentage ($\approx 8\%$) of manual auditing, (iv) Human-bias free audit, (v) Explainability (XAI) in the process of audit outcome.

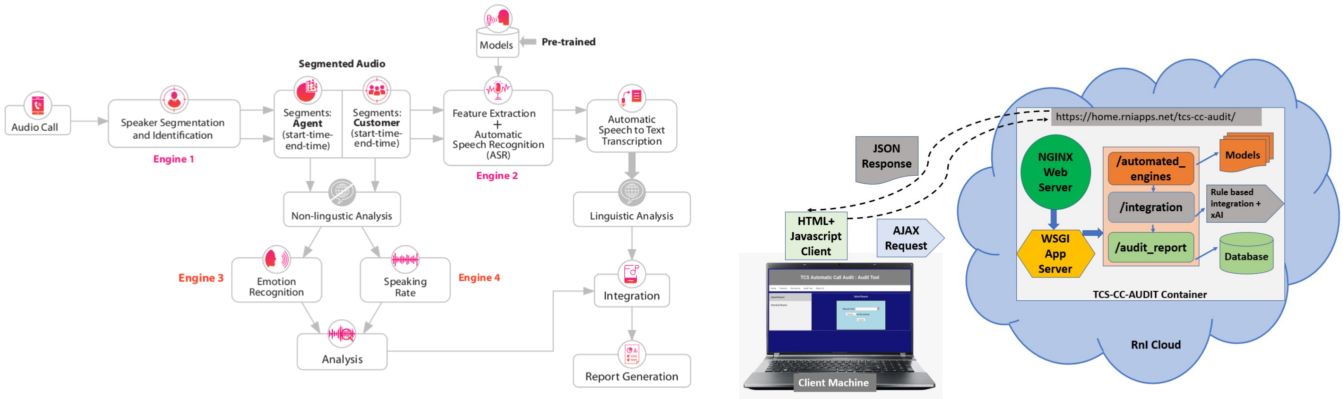
II. Automatic Call Audit

A. System Architecture

Automatic audit of call-center conversation requires several mainstream, cutting edge audio processing technologies to act together, some of them in sequence and some in parallel. For audit automation, we integrated four state-of-the-art speech engines that are researched, engineered and build in-house (see Figure 1(a)).

- Engine#1 (Speaker Segmentation and Identification): This engine is build using Kaldi's callhome diarization setup, which extracts x-vector to decide who spoke when [10]. Given a conversation, it segments (start_time, end_time) the audio according to the speaker identity speaker_ID (Speaker1 or Speaker2).
- Engine#2 (Automatic Speech to Text): This engine uses the ASR model trained using Librispeech corpus of 960 hours of speech, which is further augmented with noise and reverberation [11]. For each segment of the conversation, ASR engine converts the speech into corresponding text, to be used for linguistic analysis.
- Engine#3 (Emotion Recognition): This engine is used to recognize emotion in spontaneous conversation. Four standard emotions (anger, happy, neutral, sad), which are mostly observed in call-center scenario, are considered here. The models are created using both publicly available emotional speech and our own call-center conversation [12], [13] data.
- Engine#4 (Speaking Rate): This engine computes the speaking rate that is measured as number of words per minute (wpm) [14].

It is to be noted that the above described four engines are used without target environment specific (i.e. call-center scenario) tuning, except the language model (LM) in ASR that is tuned with call-center conversation specific key-words and key-phrases. The input to Engine#1 is an audio call, which labels each segment with speaker_ID, followed by identifying each speaker's voice as agent or customer. Once we have the output from Engine#1, remaining three engines provides transcriptions (Engine#2), emotional state (Engine#3) and speaking rate (Engine#4) of the conversational speech. Thereafter, the output of speech engines are fed to the integration module, and in order to design rules that replicate human-level reasoning for several high-level attributes, we use the deep analysis of the manual-audit process. Thus, attribute-specific set of rules are used to evaluate each parameter. The automated audit system also provides an explainable



(a) System architecture

(b) Platform design

Fig. 1. Framework : Automatic Call-center Call Audit

Algorithm 1 P3 : Paraphrasing user's concern (XAI)

```
Input: $call\_segment$ : $start$, $end$ as a % of total call duration
$A$, $C$ : Agent's and Customer's diarized speech
$kws$ : $[k_1, k_2, \cdots, k_n]$
Output: $Yes/No$ : presence/absence
condition1 : continuous (longest) $A$ preceded by $C$
condition2 : $A_{dur} > C_{dur}$
condition3 : presence of $kws$ in $A$
Parameter Evaluation:
if (condition1 or condition2 or condition3) then
    return YES {Paraphrases}
else
    return NO {Does not Paraphrases}
end if
```

TABLE I
MANUAL VS AUTOMATED: PARAMETER WISE AGREEMENT (%)

ID	**Attributes**	Phase#1	Phase#2	Phase#3
P1	Greeting & User Information	67.5	98.75	98.75
P2	Identity Verification	87.5	100.00	100.00
P3	Paraphrasing user's concern	51.25	65.00	95.00
P4	Investigation - Probing	78.75	92.50	98.75
P5	Diagnose & Troubleshooting	75.00	91.25	95.00
P6	Easy to understand and neutral accent	65.00	67.50	86.25
P7	Being warm and friendly	96.25	97.50	97.50
P8	Hold Procedure	51.25	72.50	72.50
P9	Professionalism	90.00	91.25	91.25
P10	Offering additional assistance	76.25	85.00	85.00
P11	Providing Ticket details to the user	55.00	67.50	88.75
P12	Summarize and close the call	70.00	92.50	92.50
P13	Encouraged the customer to provide feedback	50.00	57.50	95.00
P14	Going an extra mile	21.25	25.00	91.25
P15	User Delight	30.00	57.50	96.25
	Average Performance	**64.33**	**77.42**	**92.25**

insight alongside the evaluation outcome of each attribute. We show, as an example, an integration algorithm for one of the 15 parameters in 1, and all the integration logic is implemented in python.

B. Audit Platform Design

Figure 1(b) represents the web based platform design of our audit system which was deployed on an intranet local cloud. It allows auditors to upload recorded conversation in bulk amount, and in a single request. Also, the interface allows simultaneous access by multiple users. On call conversation upload, the system first estimates the expected time required to generate the audit report of the current request (based on the current server load and the predicted time to audit the uploaded files) and notifies the user when to expect the audit report. Meanwhile, the request is passed to backend Flask APIs (i.e. automated audit system with all engines as discussed in section II-A) for the audit report generation. Thereafter, user can access the same interface and download the call quality audit report.

III. SYSTEM PERFORMANCE

We used real service desk calls from a pharma industry for evaluating the performance of our audit system. As mentioned earlier, all the engines used without environment specific tuning. The automated system presented here was build through several development phases. Table I represents the attribute wise agreement (in %) achieved in 3 different phases. In the initial phase Phase#1, we build the integration logic (see Section II-A) using the knowledge assimilated from the manual audit on 40 calls. While testing on 80 calls (different from 40 calls), we achieved an agreement of 64.33% between the automated and manual audit process. We observed word misrecognition in ASR due to the use of general purpose LM. Therefore in Phase#2, we tuned LM in Engine#2 with the call-center specific kws [15], and for that we combined the 3-gram LM trained using call-center specific keywords with the Librispeech LM. This resulted in reduction of errors in automatic transcription that in turn improved the kws detection, and we achieve 77.42% agreement between automated and manual audit process. This resulted in an absolute improvement of 13.09% with respect to Phase#1. Improving further, we upgraded the integration module in Phase#3, which can mimic the human-level reasoning more efficiently in the evaluation of high-level attributes. Towards this, we analyzed the manual audit process in more details, and then updated the integration logic. The system performance improved to 92.25% which is an absolute improvement of 14.83% over Phase#2. Apart from the attribute wise evaluation, we designed the audit system to provide explanation for each attribute evaluation. The overall time complexity of our proposed system is $0.6\times$ times of the total call duration which is far less as compared to the time taken by manual audit process.

IV. CONCLUSION

In this work, we demonstrate an automatic system for auditing call-center conversation. The developed system solved the problem related to auditing call-center audio conversation which was hindered by the limited amount of data that could be processed and human bias in manual audit. Automating the whole audit process has increased the reliability, reduced the inconsistency due to human bias, improved customer satisfaction and importantly made the audit process faster.

References

[1] Shivam Mehta, Aarohi Mahajan, Sneha Chitale, and Dhananjay Raut, “Birdeview: Advance version of call monitoring system by using mining techniques,” in *2020 4th International Conference on Trends in Electronics and Informatics (ICOEI)(48184)*. IEEE, 2020, pp. 561–566.

[2] Gilad Mishne, David Carmel, Ron Hoory, Alexey Roytman, and Aya Soffer, “Automatic analysis of call-center conversations,” in *Proceedings of the 14th ACM international conference on Information and knowledge management*, 2005, pp. 453–459.

[3] Martine Garnier-Rizet, Gilles Adda, Frederik Cailliau, Jean-Luc Gauvain, Sylvie Guillemin-Lanne, Lori Lamel, Stephan Vanni, Claire Waast-Richard, et al., “Callsurf: Automatic transcription, indexing and structuration of call center conversational speech for knowledge extraction and query by content.,” in *LREC*, 2008.

[4] Arun Pande and Sunil Kumar Kopparapu, “System for conversation quality monitoring of call center conversation and a method thereof,” 2014, US Patent 8761376B2.

[5] Woosung Kim, “Online call quality monitoring for automating agent-based call centers,” in *Eighth Annual Conference of the International Speech Communication Association*, 2007.

[6] Vincenzo Pallotta and Rodolfo Delmonte, “Interaction mining: the new frontier of customer interaction analytics,” in *New Challenges in Distributed Information Filtering and Retrieval*, pp. 91–111. Springer, 2013.

[7] Sunil Kumar Kopparapu, *Non-Linguistic Analysis of Call Center Conversations*, Springer, 2015.

[8] Betül Karakus and Galip Aydin, “Call center performance evaluation using big data analytics,” in *2016 International Symposium on Networks, Computers and Communications (ISNCC)*. IEEE, 2016, pp. 1–6.

[9] Geoffrey Zweig, Olivier Siohan, George Saon, Bhuvana Ramabhadran, Daniel Povey, Lidia Mangu, and Brian Kingsbury, “Automated quality monitoring in the call center with asr and maximum entropy,” in *2006 IEEE International Conference on Acoustics Speech and Signal Processing Proceedings*. IEEE, 2006, vol. 1, pp. I–I.

[10] “Callhome diarization xvector model,” https://kaldi-asr.org/models/m6.

[11] Meet H Soni, Sonal Joshi, and Ashish Panda, “Generative noise modeling and channel simulation for robust speech recognition in unseen conditions.,” in *INTERSPEECH*, 2019, pp. 441–445.

[12] Upasana Tiwari, Meet Soni, Rupayan Chakraborty, Ashish Panda, and Sunil Kumar Kopparapu, “Multi-conditioning and data augmentation using generative noise model for speech emotion recognition in noisy conditions,” in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 7194–7198.

[13] Rupayan Chakraborty, Meghna Pandharipande, and Sunil Kumar Kopparapu, *Analyzing Emotion in Spontaneous Speech*, Springer, 2017.

[14] Meghna Abhishek Pandharipande and Sunil Kumar Kopparapu, “Real time speaking rate monitoring system,” in *2011 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC)*. IEEE, 2011, pp. 1–4.

[15] “Improving domain-specific transcription accuracy with custom language models,” https://docs.aws.amazon.com/transcribe/latest/dg/custom-language-models.html.