A Language Independent Approach to Identify Problematic Conversations in Call Centers

Meghna Abhishek Pandharipande[1] and **Sunil Kumar Kopparapu**[2], Non-members

ABSTRACT

Voice based call centers enable customers query for information by speaking to human agents. Most often these call conversations are recorded by call centers with the intent of trying to identify things that can help improve the performance of the call center to serve the customer better. Today the recorded conversations are analyzed by humans by listening to call conversations, which is both time consuming, fatigue prone and not very accurate. Additionally, humans are able to analyze only a small percentage of the total calls because of economics. In this paper which is based on [1], we propose a visual method to identify problem calls quickly. The idea is to sieve through all the calls and identify problem calls, these calls can then be further analyzed by human. We first model call conversations as a directed graph and then identify a directed graph structure associated with a normal call. All call conversations that do not have the structure of a normal call are then classified as being abnormal. We use the speaking rate feature to model call conversation because it can spot potential problem calls. We have experimented on real call center conversations acquired from different call centers and the results are encouraging.

Keywords: Call Conversations, Speaking Rate, Speech Analytics, Modeling, Visual analysis of Speech

1. INTRODUCTION

Call center businesses record customer telephonic interactions to extract valuable insight into products, strategy, services, process and operational issues. Speech Analytics (SA) is the method of automatically analyzing recorded calls to extract useful and usable information. An accurate analysis of the call center conversations shed light on some very crucial usable information that would otherwise be lost [2], [3]. Analyzing conversations is an expensive task if done manually in addition to being not comprehensive. Typically, only a very small fraction of all the recorded call conversations are carefully heard by human supervisor; further the entire duration of the conversation is heard, to determine if the call is a problematic call or not. Clearly, there is a large portion of the recorded calls that are not part of the calls listened to by the supervisors and hence more often than not the problematic calls might not be part of the human analysis.

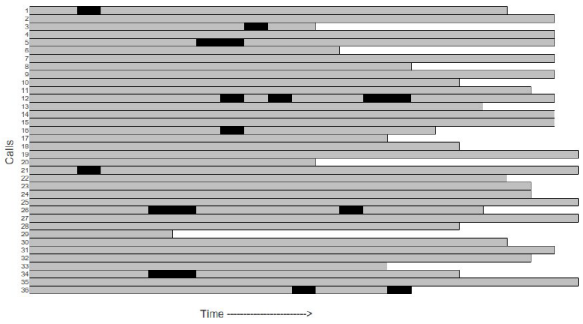

Fig.1: *Call Conversation with problems. The dark portions represent regions of problem in the call.*

The practical difficulty in identifying problematic calls manually is depicted in Fig. 1. The call duration is shown on the x-axis while different calls are depicted along the y-axis. As seen in Fig. 1 the calls are of different duration and the gray color depicts the actual length of each call. The black color within a call displays the location of a problem situation during the call. Clearly the location of problem situation is arbitrary and typically the duration of the problem situation is also very small.

Let us assume that there are N calls that the supervisor has to flag as being either normal or problematic. Let d_i denote the duration (in seconds) of the i^{th} call. For a supervisor to actually identify all the calls with a problem, he would have to listen to $\sum_{i=1}^{N} d_i$ seconds of conversation which in general is impractical in most situations. Today, supervisors randomly select a small set of calls, say some $k\%$ of N, namely,

$$M = \frac{kN}{100},$$

where $M \ll N$. For each of these M calls ($m = 1, 2, \ldots, M$), the supervisors would randomly select a point in time to listen to the call, say l^m_{start} and listen

Manuscript received on November 22, 2012.

[1,2] The authors are with TCS Innovation Labs - Mumbai Tata Consultancy Services Limited Yantra Park, Thane (West), India - 400601. E-mail: meghna.pandharipande@tcs.com and sunilkumar.kopparapu@tcs.com

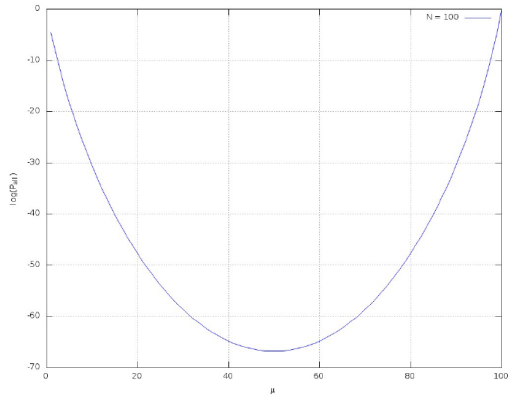

Fig.2: *Log probability of identifying all the problematic calls.*

till l^m_{stop}, namely for a duration of

$$L_m = l^m_{stop} - l^m_{start},$$

This reduces his listening time to $\sum_{m=1}^{M} L_m$ which is morepractical and can be handled by a human supervisor.

Note 1: Observe that $\sum_{m=1}^{M} L_m \ll \sum_{i=1}^{N} d_i$ because $M \ll N$ and $L_m \ll d_i$.

Suppose there are μ out of N calls that are problematic. The probability (P_{all}) that the supervisor exactly chooses those μ calls for listening is

$$P_{all} = \frac{1}{{}^{N}C_{\mu}} = \frac{(N-\mu)!\mu!}{N!} = \frac{\mu!}{\Pi_{l=0}^{\mu-1}(N-l)} \quad (1)$$

Fig. 2 shows the log probability ($log(P_{all})$) of identifying all the problematic calls correctly, from the set of N calls, as a function of the number of calls (μ) being actually problematic. Clearly if the number of problematic and normal calls are equally probable then the probability of identifying all the problematic calls exactly is low (represented by the trough in Fig. 2). Generally the number of problematic calls is about 5% to 15% of the total number of calls, while the probability is better than when the problematic calls is close to 50% of the total number of calls, it is still very small. For example, if $N = 100$ and $\mu = 10$ then using (1) we get $P_{all} = 5{:}7769 \times 10^{-14}$, which is very small. For this reason the probability of a supervisor being able to *pick up* all the problematic calls is very small; as a result most problematic calls go unnoticed for any corrective action to be taken.

However with the advent of automatic speech recognition (ASR) technology, the task of identifying problematic calls reduces to one of automatic transcription of telephone calls followed by analyzing the words or phrases in the text transcriptions [4]. However, the process of converting audio conversations into transcribed text is not very accurate, typically the recognition accuracies are around 50-60% even if one uses the state of the art speech recognition technology. The recognition accuracies further deteriorate when there is no readily available ASR for the language spoken in the conversation [5]. In many cases transcription of the audio conversation is not sufficient to identify a problem situation in a call conversation, because the problem in the call might not translate into meaningful transcribed text. For example, */thank you/* spoken in a sarcastic tone might not suggest a problem in the call because the phrase "*Thank You*" is generally associated with a satisfied customer and hence would be categorized as a non-problematic call.

In this paper, we describe a method to enable automatic identification of problematic calls without actually transcribing the audio conversations into text. We use the speaking rate [6] feature to abstract the call conversation and use directed graph to represent a call conversation. We identify a structure of the directed graph which represents a normal call. Any call conversation which does not have the structure of a normal call is then flagged as being abnormal or problematic. The main contributions of this paper are (a) Use of a non-linguistic feature, speaking rate, to represent a call conversation, and (b) modeling a call conversation as a directed graph and using this directed graph structure to identify abnormal call conversations. The rest of the paper is organized as follows. A brief literature survey in presented in Section 2, followed by modeling a call conversation in Section 3 as a directed graph. In Section 4 we describe the speaking rate feature and capture a typical call center conversation in Section 5. We present experimental results in Section 6 and conclude in Section 7.

2. RELATED WORK

Identifying problematic call conversations help the call center improve its performance in terms of customer satisfaction, customer retention, cross selling, process improvement to name a few. A spotted problematic call gives a lot of insight into possible process and people improvements that can enhance the performance of a call center. Among these benefits, the ability of the supervisor to efficiently pinpoint personalized training to the call center voice agents is very important.

There are two main approaches that call centers follow to identify problematic calls. The first approach is based on looking at the meta data associated with the call, namely if the call takes a longer time than usual to complete or if the call is transferred to a supervisor for completion, then it is flagged as being problematic. As an extension, in case when call centers are equipped with multiple interaction chan-

nels, then a complaint from a customer in the form of an email is associated with the voice call to mark it as being problematic. These approaches are not reliable as they look only at the meta data associated with the voice call and not the actual call conversation itself. Very often these meta data based approaches miss out on important problematic calls that do not leave any cues in the meta data.

The second approach is to transcribe the call conversation using an automatic speech recognition engine and then sieve through the text transcriptions to flag calls based on key words or phrases. This approach has several drawbacks, the first and the foremost is the fact that the state of the art call center conversation transcription accuracy is very noisy and erroneous and the second important factor is that there are several instances, as described earlier, when just analyzing the text transcription does not yield clues that can be associated with a call being labeled as being normal or problematic.

Hironori [7] describes a method to identify important segments from transcribed textual records of conversations between customers and agents. They look for changes in the accuracy of a categorizer designed to separate different business outcomes.

Gilad [8] describes a system that automatically transcribes calls using a speech recognition engine. The domain specific importance of the conversation fragments is identified based on the divergence of corpus statistics. This is used to analyze the content of the call conversation. They further use information retrieval approaches on the transcribed text to provide knowledge mining tools for both call-center agent and for administrators of the center. The system developed in [8] helps in gaining insight hidden in the recorded calls, which can help reduce cost of operation and improve products, processes. This enables making quality monitoring more effective by routing calls about key business issues to supervisor for review affecting the overall customer experience.

Vincenzo [9] provides a solution for pragmatic analysis of call center conversations in order to provide useful insights for enhancing Call Center Analytics to a level that will enable new metrics and Key Performance Indicators (KPIs) beyond the standard approach. These metrics rely on understanding the dynamics of conversations by highlighting the way participants discuss about topics. By this, they claim, one can detect situations that are simply impossible to detect with standard approaches such as controversial topics, customer-oriented behaviors and also predict customer ratings.

Most of the work described in literature is carried out on the transcribed call conversation, which means we have to depend on the not so accurate ASR for audio transcribed text. However there are several languages, generally called the resource deficit language, that do not have even a basic speech to text conversion engine. Even for languages, like English, where a ASR exists, the typically speech to text conversion yield only 50-60% recognition accuracies.

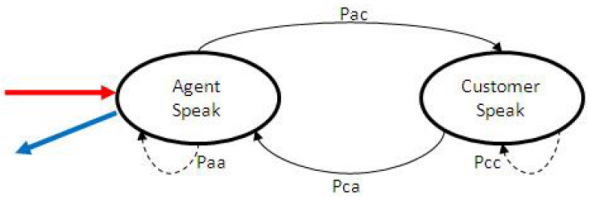

Fig.3: *Directed Graph representation of a Call Center Conversation.*

In this paper, we propose a method to flag problematic calls by neither looking at the meta data associated with the call or requiring to go through the poor accuracy speech to text transcription. We use

1) the ability to distinguish the call conversation into segments spoken by the agent and those spoken by the customer and
2) non-linguistic feature associated with the call conversation, namely, *speaking rate*

to identify and flag problematic calls. Clearly this approach is independent of the language of communication in the call conversation.

3. MODELING CALL CONVERSATIONS

A typical call center conversation between the agent in the call center and the customer is a sequence of speech segments spoken by either the agent or the customer. Generally the call is initiated by the customer and lands on the interactive voice response (IVR) system. The call is then routed to one of the several agents who are free at that time to receive a call; the voice agent then starts the conversation. A call center conversation can be represented as a directed graph as shown in Fig. 3. The nodes represents the person (agent or customer) who is speaking and edges represent the transition between the speakers.

Typically, at the beginning of a call, the agent starts with a welcome message and it is the agent who ends the call with a good bye or thank you message. As seen in Fig. 3 the red arrow into the AGENT SPEAK node shows the start of the conversation, while the blue arrow, going out of the AGENT SPEAK node shows the end of the conversation. The rest of the conversation is spoken by the customer (CUSTOMER SPEAK) or the agent (AGENT SPEAK) and is interlaced, this is shown by continuous black lines in Fig. 3. The dashed lines in Fig. 3 represent the agent (customer) speaking continuously with possible pauses without allowing the customer (agent) to speak. The edge weights P_{cc}, P_{ca}, P_{ac} and P_{aa} represent the probability that the customer continues to speak with possible pauses, probability that the agent speaks after the customer,

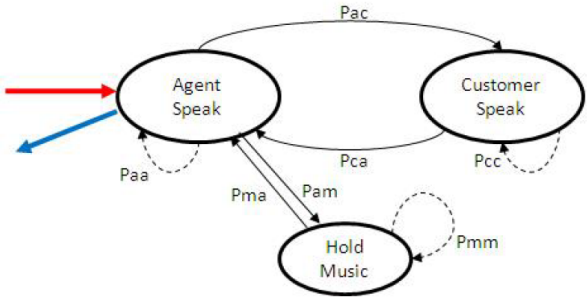

Fig.4: *Directed Graph representation of a Realistic Call Center Conversation.*

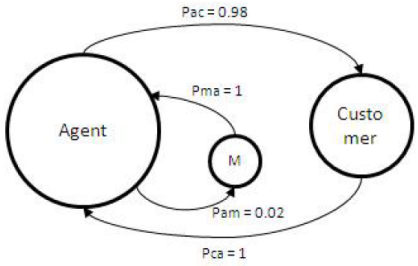

Fig.5: *Directed Graph representation of an actual normal call.*

probability that the customer speaks after the agent and probability that the agent continues to speak with possible pauses respectively.

In some sense Fig. 3 gives an abstract view of a typical call center conversation. The size of the node captures the total talk time of that speaker during the call. For example, a large AGENT SPEAK compared to CUSTOMER SPEAK node means that the agent was conversing more during the call. Additionally, the edges in the graph determines and captures the nature of the conversation. For example a missing edge out of the AGENT SPEAK node at the end of the conversation is likely to be a problematic call or if $P_{cc} \ll P_{ca}$ then the customer is speaking more and, possibly, not allowing the agent to speak or respond to his problem; in majority of the cases this might translate to a problematic call. During a normal interaction between the customer and the agent one expects $P_{ac} > P_{aa}$ and $P_{ca} > P_{cc}$ where the agent speaks and also allows the customer to speak and vice-versa.

However a realistic call conversation has *holdtime*, the time during the conversation, when neither the agent nor the customer is speaking; this happens when the agent is fetching the information sought by the customer from an IT system, so there is another node, namely, HOLD MUSIC associated with the conversation. In a typical conversation the HOLD MUSIC state is entered from the AGENT SPEAK state and exits to AGENT SPEAK state. Clearly if the conversation stays longer in the HOLD MUSIC state (large P_{mm}, see Fig. 4.), then it can be assumed that the agent is not performing efficiently because he is putting the customer on wait for a longer duration or the customer has a complex query that the agent is not being able to address. Along similar lines, if P_{am} is larger than P_{aa} or P_{ac} then it is very likely that the agent is putting the customer on hold a large number of times indicating that the agent is unable to resolve the customer query. These situations are reasonably good indicators of a problem situation in a call center conversation. Clearly this mode of modeling a call center conversation enables one to understand certain aspects of the call in terms of performance of the agent. Understandably both (a) the probabilities P_{**} and (b) the size of the nodes, AGENT SPEAK, HOLD MUSIC and CUSTOMER SPEAK are computed on the complete call and hence gives an overall picture of the call.

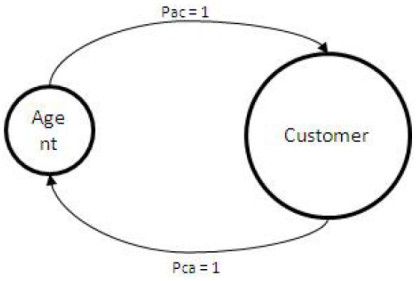

Fig.6: *Directed Graph representation on an actual abnormal Call.*

Note 2: The probabilities P_{**} and the size of the nodes, in a call conversation are computed as shown in Appendix A.

Note 3: The probabilities P_{**} do not in themselves help in identifying the portions of the call that might not be normal.

As will be shown later, the directed graph representation of the call conversation has a different structure which depends on the type of transaction happening during the call.

To check the use of the directed graph to identify abnormal calls we randomly selected 75 actual call center conversations that we had access to from call centers. Fig. 5 shows a typical directed graph of a normal call while Fig. 6 shows the directed graph of an abnormal or a problematic call. As a first step the calls were automatically segmented into voice and music [10]. Further these voice segments were segmented into sections spoken by agent and customer using an automated method described in [11]. Subsequently, we found the number of transitions from one node (here node refers to AGENT SPEAK, CUSTOMER SPEAK and HOLD MUSIC) to another and also calculated the duration of speech in a particular node (as described in Appendix A). It was found that in a normal call the probability of agent talking to customer P_{ac} is lesser than the probability of customer talking after an agent has spoken P_{ca}

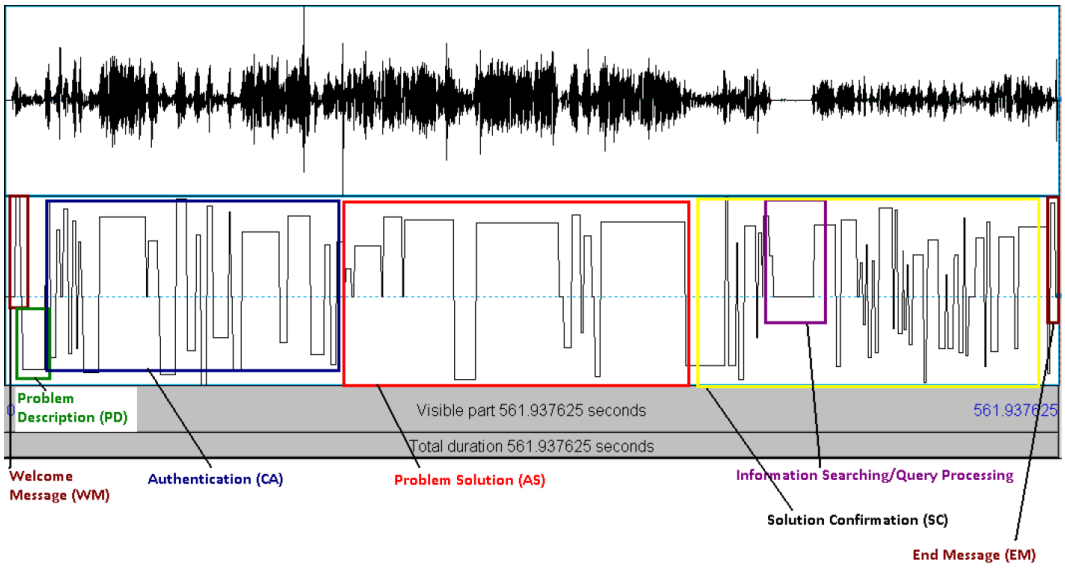

Fig.7: *Different transactions in a Call Conversation and the associated Speaking Rate Pattern.*

($P_{ac} < P_{ca}$). Also the probability of an agent putting the customer on hold P_{am} is much lesser than probability of agent talking to customer P_{ac}, namely, $P_{am} \ll P_{ac}$. Note that the size of the AGENT SPEAK node (P_{aa}) is much larger than the node corresponding to CUSTOMER SPEAK (captured by P_{cc}) meaning the agent is talking for more duration than the customer. This in general symbolizes a typical normal call. While in an abnormal call see Fig. 6 the size of the node corresponding to the CUSTOMER SPEAK is much bigger than the size of the node corresponding to AGENT SPEAK typically indicating a problematic call.

However, this analysis based on directed graph is not sufficient to identify the actual location of the abnormality, if it exists, in a call. We need to model the call at a better time resolution, we discuss this next where we use the speaking rate as an additional non-linguistic feature to represent the call conversation.

4. SPEAKING RATE FEATURE

Speaking rate is measured as number of spoken words per minute (WPM). While there are several approaches to identify the speaking rate, one of the well known method of measuring speaking rate involves identifying the syllables in the spoken speech to compute the speaking rate. An algorithm to detect a syllable reliably in a spoken speech has been described in [12]. The syllable detection is based on the intensity (loudness) and the voicedness of the spoken speech. The pauses in the spoken speech are also identified so that the number of occurrences of the syllables per unit spoken time gives an accurate measure of the speaking rate. In all our experiments we used the algorithm in [12] to determine the speaking rate. Once the speaking rate is computed in terms of

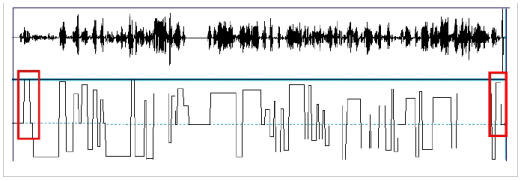

Fig.8: *Speaking Rate is high at the start and end of the call.*

number of syllables per second (sps), we can compute the speaking rate in words per minute (wpm) using a conversion factor of $\gamma = 1{:}5$ as suggested by Yaruss [13], namely,

$$SR_{wpm} = \gamma \times SR_{sps} \times 60, \tag{2}$$

where SR_{wpm} is the speaking rate in words per minute, SR_{sps} is the number of syllables per second and γ is the conversion factor between syllable and word.

Note 4: The factor γ is dependent on the language and this choice of $\gamma = 1{:}5$ is applicable to conversational English. For analysis of call conversation we used the speaking rate feature. The speaking rate was computed on a segment of the call conversation spoken by the same person (agent or customer) in one stretch. As seen in Fig. 7, a typical call conversation can be segmented into portions of speech spoken by the agent (AGENT SPEAK), that spoken by the customer (CUSTOMER SPEAK) and the hold music (HOLD MUSIC). Each of this segment which is not hold music, can be analyzed to estimate the speaking rate (SR_{wpm}), the speaking rate during hold music segment is assumed to be zero. The speaking rate of the agent is shown on the positive y-axis and that of the customer is shown on the negative y-axis. The

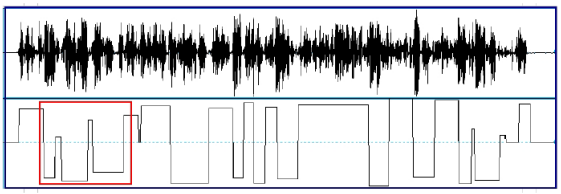

Fig.9: *Problem Description by Customer.*

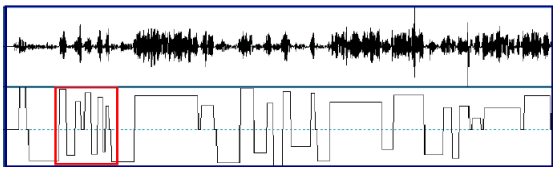

Fig.10: *Authentication Process.*

amplitude represents the speaking rate; larger amplitude represents a higher speaking rate and vice-versa. A typical speaking rate pattern of a complete call conversation of 598:7 seconds duration is shown in Fig. 7.

Note 5: Clearly, since we are representing the speaking rate on the negative y-axis for the customer, one needs to consider the absolute value of the speaking rate so a higher absolute value would represent faster speaking rate.

5. SPEAKING RATE PATTERNS IN CALL CONVERSATIONS

There are several patterns that one can observe in the call conversations that are typical of a conversation between the customer and the agent. Speaking Rate of the agent is high at the start of the conversation as well as at the end of a call conversation as seen in Fig.8.

This is to be expected as very often the agent is either reading a predrafted script or has spoken the paragraph so many times that it becomes his second

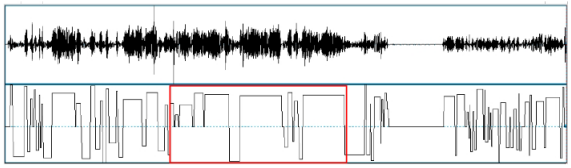

Fig.11: *Solution being provided by agent.*

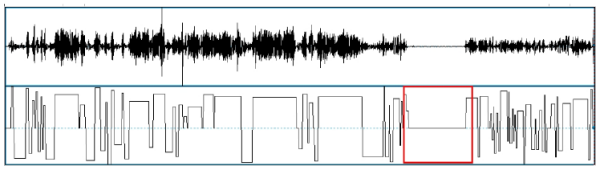

Fig.12: *Agent searching for Information.*

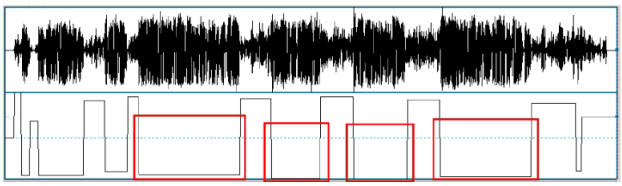

Fig.13: *Abnormal Call: Customer not allowing agent to speak.*

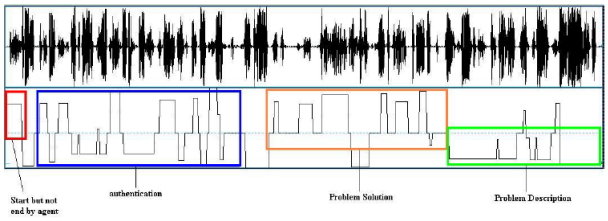

Fig.14: *Abnormal Call: Call ends abruptly no /Thanks message/ from agent.*

nature. For example: */very good evening, welcome to < company > care this is < name > how may I assist you/* at the beginning of the call or */thank you for calling < company > care have a nice day/* at the end of the call.

A customer describing the problem shows up as a pattern shown in Fig. 9. The pattern has more or less a constant speaking rate with intermittent agent speaks of very short duration. Fig. 10. shows the authentication procedure. Typically, the agent makes sure that he is indeed speaking to the person whom he is supposed to speak. Asking for name or contact number or card details to verify and authenticate the customer. Agent providing a solution in response to the customer query has a pattern shown in Fig. 11. Typically the agent speaks for a length of time with small pauses in between; this is an indicative pattern of the agent providing a verbal solution to the customer query and an agent searching for some information has a pattern shown in Fig. 12.

We can observe that the speaking rate feature displays unique patterns which capture certain aspects of the call conversation. Additionally using speaking rate feature we can very clearly identify all those instances in the conversation where the customers speaking rate is continuously high. Fig. 13, for example, captures an instance in a conversation where the customer is speaking continuously with a high speaking rate, this pattern is an indication of the customer

Table 1: *PATTERNS IN A NORMAL CALL.*

Transaction	Pattern (SR)	Who
Welcome Message (WM)	High	Agent
Problem Description (PD)	Uniform	Customer
Authentication (CA)	Low	Agent-Customer
Agent Solution (AS)	Uniform	Agent
Solution Confirmation (SC)	Uniform	Agent-Customer
End Message (EM)	High	Agent

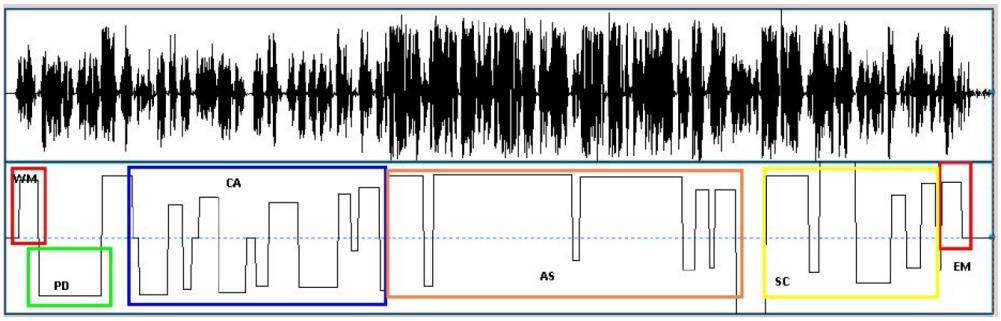

Fig.15: *Typical Speaking Rate Pattern of a Call Conversation.*

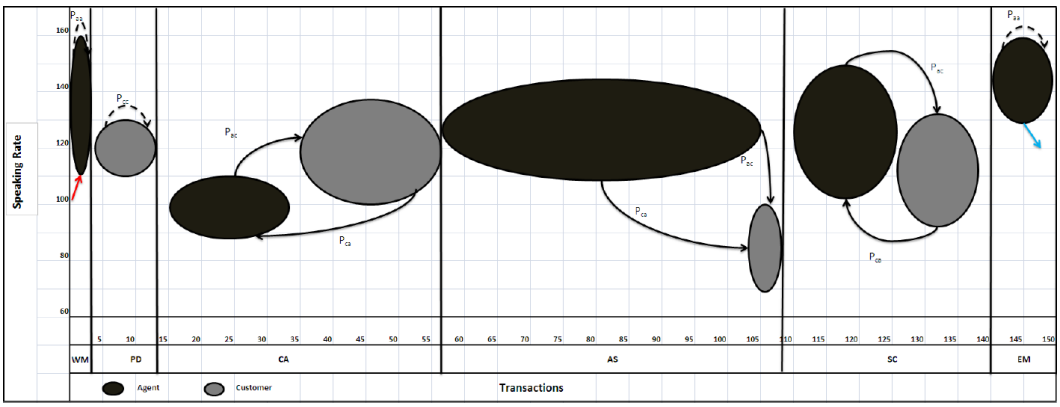

Fig.16: *Structure of the directed graph in each transaction.*

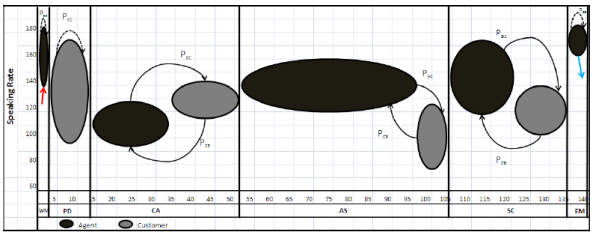

Fig.17: *Directed graph in each transaction for a different call.*

not being happy for some reason. Also conversation ending abruptly with no */Thank you/* message from the agent (Fig. 14) is an indication of an upset customer who hung off the call midway.

The observed patterns in the speaking rate feature can be associated with a typical transaction within the call conversation. Typical transactions in a conversation could be

1) $[WM]$ Welcome Message by the agent
2) $[PD]$ Problem Description by customer
3) $[CA]$ Customer Authentication,
4) $[AS]$ Agent providing a Solution
5) $[SC]$ Agent Confirming customer understanding the solution, solution confirmation.
6) $[OP]$ Pauses, Music
7) $[EM]$ End Message by the agent.

Fig. 15 (like Fig. 7) captures a complete call center conversation. The red box at the beginning is the $[WM]$ component and indicates the start and the red box at the end of call indicates the component $[EM]$. The green box indicates description of problem by customer which is the $[PD]$ component while the blue box indicates the authentication process $[CA]$, where the agent makes sure that he is indeed speaking to the person he should be speaking to. The orange box indicates a typical solution being provided by the agent to customer $[AS]$. The yellow box indicates confirmation of the solution provided by the agent and affirmation by the customer to the solution $[SC]$. Table I captures this in a nutshell. Each of these transactions within the call conversation can be represented as a directed graph, consisting of AGENT SPEAK, CUSTOMER SPEAK and HOLD MUSIC. Figs. 16 and 17 capture the structure of the directed graph in each of these transactions for two different call conversations. The y-axis shows the speaking rate and the x-axis is an indication of the time in the conversation. The size or the area of the node captures the total speaking time during the transaction and the ellipticity of the node shows the variation in the speaking rate during the transaction. Clearly, in the transaction corresponding to [WM], only the agent is speaking (larger AGENT SPEAK) and he is speaking fast (location of the AGENT SPEAK node is higher up along the y-axis). As seen in Fig. 16, a normal call has the following characteristics, namely,

1) A normal call has a very specific pattern in terms

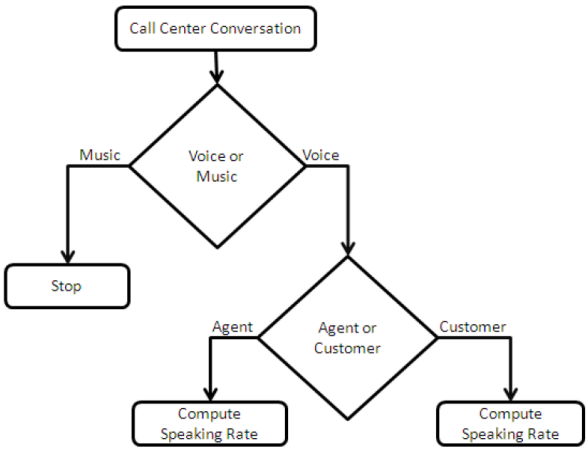

Fig.18: *Computing the SR of agent and customer..*

of transactions components during the length of the conversation, and

2) A normal call has typical components which are organized in a particular time sequence.
3) Normal call conversations are structurally similar as seen in Fig. 16 and Fig. 17.

In our experiments, we used the structure defined by Fig. 15 and Fig. 16 or captured in Table I to identify a normal call. We compared test call conversation with this reference model to identify how close the call was to a normal call. Any deviation from the normal structure was flagged as being abnormal.

6. EXPERIMENTAL RESULTS

For the purpose of experiments we analyzed a set of 100 call center conversations obtained from three different call center catering to Insurance and Telecom domain, while the calls associated with the insurance domain were in English language they were from two different geographies; the telecom domain related calls were in a mix of Hindi and English language. All these calls were actual customer-agent calls and were recorded at the call center sampled at 8 kHz and each sample was represented by 16 bits and stored in the *.wav* format.

All these 100 calls were carefully listened to by atleast a supervisor in the call center and each was marked as being abnormal or normal. Of these 100 calls 10 calls were marked abnormal. For example, one of the calls, marked as abnormal, the customer had put down the phone without allowing the agent to even speak his typical thank you message. The other 90 calls were marked as being normal by the call center supervisor. For a given conversation (see Fig. 18), we

- Automatically segmented the conversation into segments corresponding to voice and music using the method [10],
- and then all the segments that were marked as voice were further automatically segmented into spoken by agent or spoken by customer refer [11].
- For each of the voice segment, we automatically computed the speaking rate of the segment using the method mentioned in [6].

We marked the conversations, manually, using the speaking rate feature to determine the type of transaction, namely, one of $[WM], [PD], [CA], [AS], [SC], [OP], [EM]$ without actually hearing to the audio conversation (using Table I). We selected, at random, a few call conversations to check the correctness of the marking of the transaction by manually listening to the audio segments, and found that we were able to mark with more than 90 % accuracy. Now each call conversation was first segmented into voice and music [10] and then all the segments that were marked as voice we segmented into spoken by agent or spoken by customer [11]. We then constructed the directed graph for each transaction in the call conversation by first labeling the segment with one of the labels ($[WM], [PD], [CA], [AS], [SC], [OP], [EM]$) along with the time taken for that transaction. Any call that (a) missed one or more labels, or (b) a certain label that had an unusual duration was marked as being abnormal. This process yielded 90 % results in the sense that we were unable to mark just 1 abnormal call in the 10 abnormal calls in our dataset. However, as many as 7 normal calls were marked as being abnormal because we had falsely mislabeled the conversation with transaction labels based on the observed pattern.

Note 6: We are in the process of identifying pattern matching techniques that can be used to reliably mark transaction label in the conversation.

Note 7: All the normal conversation had a directed graph representation that resembled Fig. 16.

7. CONCLUSION

There is a rich source of information begging to be exploited in the customer-agent voice conversations which can enhance the customer satisfaction index and other performance metrics of a voice based call center. Additionally, information derived from the call conversation can be used to identify personalized training needs of the agent as well as quickly address area of concerns raised by specific customers. Voice based call centers are either counting on individual supervisors to manually analyze a small sample of the recorded call conversation, or arent performing any analysis at all.

However there is a growing awareness, within the call center industry, to exploit the information hidden in the call conversations which has led to voice based call centers adopting automatic methods to analyze the conversations. The tools that have been adopted to enable speech analytics, have been restricted to the process of first converting the audio conversation into text using a speech recognition engine, followed by text analytics. This process, though natural, is

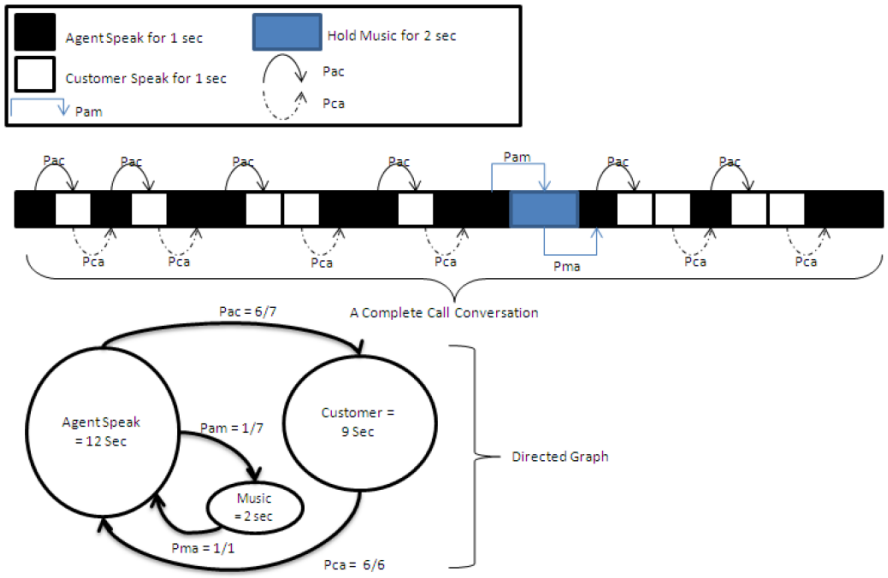

Fig.19: *Computing P_{**} from a complete call.*

not only expensive but is also erroneous. The two step process requires a speech to text conversion, which in itself returns poor recognition accuracies even with the state-of-art speech recognition engines. Additionally, since the analysis is based on text, there is a challenge in determining a */thankyou/* said with sarcasm versus a genuine */thankyou/*. This poses problem in determining an abnormal or a problematic call from a normal call conversation just based on analyzing text. In this paper, we have proposed a novel way of visually identifying an abnormal call from a normal call center conversation without actually converting the conversation to text or manually listening to the call. We proposed the use of directed graph to represent a complete call conversation between the agent and the customer, which is one of the contributions of this paper. Additionally, we proposed the use of a non-linguistic feature, namely, speaking rate to enable identification of different types of transactions within a call conversation. The identification of the speaking rate patterns and mapping them to transactions in the call conversation is another contribution of this paper. The experimental results show that it is possible to indeed identify abnormal call center conversations using this methodology with good accuracies. The main advantage of using the methodology suggested in this paper is that the proposed method is independent of the language of conversation and does not require the use of a language dependent speech recognition engine.

References

[1] M. A. Pandharipande and S. K. Kopparapu, "A novel approach to identify problematic call center conversations, in *Computer Science and Software Engineering (JCSSE), 2012 International Joint Conference on*, 30 2012-June 1, pp. 1-5.

[2] C. Bailor, "The why factor in speech analytics, *CRM Magazine*, vol. August, 2006.

[3] wikipedia. [Online]. Available: http://www.wikipedia.com/

[4] VERINT, "Speech analytics essentials for audiolog, *IMPACT360*.

[5] S. K. Kopparapu and I. Ahmed, "Enabling rapid prototyping of an existing speech solution into another language, in *14th Oriental COCOSDA Conference*, Oct 2011.

[6] M. Pandharipande and S. Kopparapu, "Real time speaking rate monitoring system, in *Signal Processing, Communications and Computing (ICSPCC), 2011 IEEE International Conference on*, sept. 2011, pp. 1-4.

[7] H. Takeuchi, L. V. Subramaniam, T. Nasukawa, and S. Roy, "Automatic identification of important segments and expressions for mining of business-oriented conversations at contact centers, in *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*, 2007.

[8] G. Mishne, D. Carmel, and R. Hoory, "Automatic analysis of call center conversations, in *ACM international conference on Information and knowledge management*, 2005.

[9] V. Pallotta, R. Delmonte, L. Vrieling, and D. Walker, "Interaction mining: the new frontier of call center analytics, in *5th International Workshop on New Challenges in Distributed Information Filtering and Retrieval*, 2011.

[10] S. Kopparapu, M. Pandharipande, and G. Sita, "Music and vocal separation using multiband

modulation based features, in *Industrial Electronics Applications (ISIEA), 2010 IEEE Symposium on*, oct. 2010, pp. 733-737.

[11] S. Kopparapu, A. Imran, and G. Sita, "A two pass algorithm for speaker change detection, in *TENCON 2010 - 2010 IEEE Region 10 Conference*, nov. 2010, pp. 755-758.

[12] N. H. D. Jong and T. Wempe, "Praat script to detect syllable nuclei and measure speech rate automatically, *Behavior Research Methods*, vol. 41, pp. 385-390, 2009.

[13] J. S. Yaruss, "Converting between word and syllable counts in childrens conversational speech samples, *Journal of Fluency Disorders*, vol. 25, no. 4, pp. 305 316, 2000. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0094730X00000887

APPENDIX A
ESTIMATING P_{**}

It is assumed that the call conversation has been marked into segments that have been spoken by agent (AGENT SPEAK), by customer (CUSTOMER SPEAK) or was on hold (HOLD MUSIC). Fig. 19 shows the method used to compute all the probabilities P_{**} and the size of the node for a call conversation of 23 sec duration. In terms of notation, the segments marked in black represent the portion of the call where the agent is speaking (AGENT SPEAK), segments marked in white represent the portion of the call where customer is speaking (CUSTOMER SPEAK) and the segments marked in blue represent the call where the customer is on hold (HOLD MUSIC). The probability P_{ac} is the probability of a customer speaking after the agent has spoken, P_{ca} is the probability of agent speaking after the customer has spoken, P_{am} is the probability of the agent putting the customer on HOLD MUSIC and P_{ma} is the probability of the customer conversing with the agent after being on hold.

The directed graph model represents the overall picture of the complete call. The size of the node AGENT SPEAK represents the total duration of the conversation when the agent was speaking (12 seconds in Fig. 19), similarly the size of the node CUSTOMER SPEAK captures the total duration for which the customer was speaking during the entire call (9 seconds), and the size of the node HOLD MUSIC represents the total duration for which the customer was on hold (2 seconds). Clearly, in this sample call it can be visualized (from the size of the node) that the agent has spoken for the longest duration.

Similarly we can compute, automatically, the various probabilities P_{**} as follows. As seen in Fig. 19 there are a total of 6 instances when the customer spoke after the agent (marked by P_{ac} in Fig. 19) and only one instance of the music coming into existence after the agent spoke (marked as P_{am}). So, we can compute $P_{ac} = \frac{6}{7}$ and $P_{am} = \frac{1}{7}$. On the other hand, as seen in Fig. 19 there are 6 transitions from CUSTOMER SPEAK to AGENT SPEAK and no transition from CUSTOMER SPEAK to HOLD MUSIC, so we compute $P_{ca} = \frac{6}{6}$ and $P_{cm} = \frac{0}{6}$. Similarly, there is only 1 transition from HOLD MUSIC to AGENT SPEAKand no transition from HOLD MUSIC to CUSTOMER SPEAK. Using this we can compute $P_{ma} = \frac{1}{1}$ and $P_{mc} = \frac{0}{1}$.

Meghna Pandharipande received her BE in Electronics and Telecommunication in June 2002 from Amravati University. Between September 2002 to December 2003, she was a faculty member in the Department of Electronics and Telecommunication at Shah and Anchor College of Engineering, Mumbai. In 2004, she did her certification in Embedded Systems from CMC, Mumbai and then worked as a Lotus Notes developer in a startup ATS, Mumbai for about a year. Since June 2005 she has been with TCS (having first joined Cognitive Systems Research Laboratory, Tata Infotech under Prof. PVS Rao) and since 2006, she has been working as a Researcher at TCS Innovation Labs, Tata Consultancy Services, Mumbai.

Her research interest is in the areas of Speech Signal processing and has been working extensively in trying to build systems that can process all aspects of spoken speech. Some of her earlier work includes robust detection of keywords in Multilingual News broadcast videos, development of a metric to enable speech data clustering, use of non-linear audio features to separate Voice and Music. More recently she is working in the research area of non-linguistic aspects of speech processing like speaking rate and Emotion detection from speech.

Meghna is the author of several papers and has presented her work at several conferences.

Sunil Kumar Kopparapu obtained his Ph.D. degree in electrical engineering from the Indian Institute of Technology, Bombay, India, in 1997. His thesis was titled "Modular integration for low-level and high-level vision problems in a multi-resolution framework." He worked at the Automation Group, Commonwealth Scientific and Industrial Research Organization (CSIRO), Australia, between 1997 and 2000. He worked on certain practical image processing and 3D vision problems, mainly for the Australian mining industry.

Prior to joining the Cognitive Systems Research Laboratory (CSRL), Tata Infotech Limited, as a Senior Research Member, in 2001, he was associated with the Research and Development Group at Aquila Technologies Private Limited, India, as an Expert for developing virtual self line of e-commerce products. Presently, he is a senior scientist with the TCS Innovations Labs - Mumbai and is actively working in the areas of speech, script, image and natural language processing with the focus on building usable systems for mass use.

He has coauthored a book titled Bayesian Approach to Image Interpretation in addition to several publications.