Listening To Silences In Contact Center Conversations Using Textual Cues

Digvijay Ingle, Ayush Kumar, Jithendra Vepa

Observe.AI, India

digvijay.ingle@observe.ai, ayush@observe.ai, jithendra@observe.ai

Abstract

Contact center conversations often consist of silent segments, where neither the customer nor the agent is speaking. These silences if continued beyond an acceptable level can negatively impact contact center KPIs. Thus, understanding silences and defining measures to handle them better via appropriate coaching and alerting for agents is one of the key focus areas for contact centers. In this paper, we demonstrate how dialogue turns around silences could be used to understand the characteristics of silences (*expected* vs *unexpected* and *agent* vs *customer*-caused silences) via two text classification tasks. We propose a methodology to pre-train a silence-aware language model in contact center domain, called Silence-RoBERTa and demonstrate its ability to better capture the conversational characteristics around silences. Finally, we discuss the application of the above methodology in real-time and post-call settings and demonstrate its usability to reduce silences via a real-life case study.

Index Terms: contact center, language model, BERT, RoBERTa, silence

1. Introduction

Contact center conversations often comprise of silence when either agents or customers engage in some off-call work, resulting in segments with no meaningful conversation. We call these segments as *Conversational Silences*. While these silences are inevitable in natural conversations, uninformed, frequent and long silences often lead to negative impact on the business metrics in two-folds: 1) Increase in average handle time (AHT) of calls, 2) Poor customer experience.

In order to improve the operational metrics of AHT and customer experience, contact centers often aim to optimize silences in calls by focusing on long and frequent silences. However, a closer look at the conversation unveils that the scope of understanding silences is not only limited to duration but can potentially be extended to how the agent-customer interaction went on around the silence. Table 1 illustrates a few scenarios where silences naturally appear in conversations, however the silences that appear in an unexpected manner (Examples 3 and 4) or expected silences that continue for longer duration tend to be the reasons for poor customer experience. Thus, determining if a silence is *expected* or *unexpected* is one of the critical aspects from the customer experience standpoint. We believe that conversation around silence contains information indicative of characteristics of the silence. Hence, understanding characteristics of silences is not merely an audio problem but is also closely connected to understanding surrounding text.

Furthermore, it is important for contact centers to be able to define actionable strategies to deal with silences. To that effect, agents or contact centers have limited control on silences caused by customers. Thus, being able to determine the *causer of silence* would help contact centers to employ focused efforts for designing corrective actions on silences caused by agents.

Additionally, supervisors evaluate agents on their handling of silences and provide them feedback. However, these evaluations are done weeks after the onset of call which makes the process reactive rather than proactive. Thus, having a real-time feedback mechanism would allow agents to immediately course correct and ensure better adherence to protocol. Hence, we extend the definition of silence understanding to real-time setting to provide proactive alerts to agents for using correct prompts.

Thus, we propose a framework to understand silences in contact center conversations. The contributions of this work are three-fold:

1. Propose an automated method for monitoring and managing silences in contact center calls in both real-time and post-call settings.
2. Design a framework for understanding silences via two text classification tasks: 1) Type of silence - *expected* or *unexpected*, 2) Causer of silence - *agent* or *customer*.
3. Formulate a methodology for pre-training silence-aware language model in contact center domain.

2. Problem Formulation

To understand the characteristics of silence, we formulate following two Spoken Language Understanding (SLU) tasks:

1. **Silence-Type Classification**: Given a fixed context of dialogue around a silent segment, we frame a binary text classification task with *Expected Silence* and *Unexpected Silence* as the associated labels. Expected Silence denotes the category of silences where the speaker either explicitly prompts about upcoming silence or there is a mutual understanding between the parties about it (Example 1 and 2 in Table 1).
2. **Causer Identification**: Given a fixed context around a silent segment, we frame a binary text classification task to predict causer of the silence - *Customer* or *Agent*. Contact centers have limited control over silences that are caused by customers, making them relatively less actionable. Thus, the ability to predict the causer of silence would help contact centers identify agents who are frequently missing the protocol and thereby design dedicated coaching sessions for them.

3. Methodology

3.1. Pre-train Silence-Aware Language Model

While speech is a primary medium of conversation in spoken language, silences often offer contextual cues to it [1]. Typical

10.21437/Interspeech.2023-1791

Table 1: *Illustrative Examples of Silence. The color codes are as follows: a) Blue - Actual silent segment in the call, b) Pink - Prompt used to give indication about silence, c) Yellow - Instructions given by agent to customer*

S.No	Example	Characteristic
1	**Agent:** do you mind if i place you on a brief hold **Customer:** sure please go ahead SILENCE *<agent taking time to search for information>* **Agent:** thank you for holding	**Type:** Expected silence **Causer:** Agent **Explanation:** Agent explicitly informed the customer before placing on hold
2	**Agent:** go to top of your screen and click on login **Customer:** okay **Agent:** now enter the details that you just received via email SILENCE *<customer taking time to enter the details>* **Customer:** logged in	**Type:** Expected silence **Causer:** Customer **Explanation:** Agent is giving step-by-step instructions to customer and hence there is an implicit expectation that customer might take time to follow them
3	**Agent:** thank you for calling xyz logistics **Customer:** i just got a call from you all SILENCE *<abrupt silence>* **Agent:** ok what is your last name	**Type:** Unexpected silence **Causer:** Agent **Explanation:** Agent intentionally or unintentionally takes time to respond
4	**Agent:** your email id please **Customer:** abc at xyz dot com SILENCE *<agent taking time to search for information>* **Customer:** hello are you there	**Type:** Unexpected silence **Causer:** Agent **Explanation:** No prior indication given by agent regarding silence

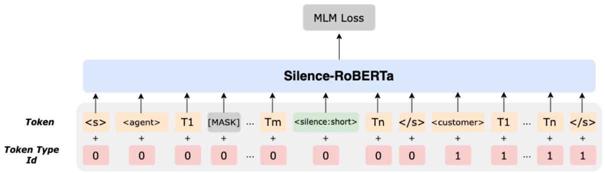

Figure 1: *Pre-Training Silence-RoBERTa Model*

SLU systems tend to rely only on text to learn conversational representations. However, we hypothesise that augmenting it with silence information provides additional context to learn better representations. For example, presence of SILENCE below helps to disambiguate if the agent is seeking additional information from the customer while looking into their account (scenario 1) or is looking into their account to confirm the order placement date mentioned by the customer (scenario 2).

1. **Scenario 1:** *let me take a look at your account SILENCE just to confirm you said you placed the order yesterday*
2. **Scenario 2:** *let me take a look at your account just to confirm you said you placed the order yesterday*

As a result of this disambiguating nature of silences, we propose a simple yet effective approach to pre-train silence-aware language model (*Silence-RoBERTa*) by including special tokens representing silences in conversations. We describe the pre-training process in detail below.

3.1.1. Data

To train the silence-aware language model, we use time-aligned ASR transcripts of approximately 2.1M English dyadic conversations between agents and customers. These conversations are sampled from a proprietary dataset[1] of contact center calls with average duration of 7 minutes and spanning across multiple lines of business like retail, e-commerce, finance, healthcare, etc. dealing with calls from service and support verticals.

3.1.2. Pre-Processing

Each of the transcripts are pre-processed at turn level. Individual turns are prepended with special tokens `<agent>` or `<customer>`, to indicate the speaker of the turn and appended with `</s>` token to mark the end of the turn. Silent segments in the transcripts are encoded based on difference in timestamps between consecutive tokens. We create 3 bins of silences based on duration - *short* (3-5 seconds), *medium* (5-10 seconds), *long* (>10 seconds). Each of these are represented by inserting special tokens `<silence:short>`, `<silence:medium>` and `<silence:long>` respectively in the transcript. For pre-training, random 15% tokens are replaced with `[MASK]` token. Finally, an additional sequence of *token type ids* is created to map individual tokens in a turn to corresponding speakers. The encoding process has been illustrated in Figure 1.

3.1.3. Pre-Training Implementation Details

We train the silence-aware language model (*Silence-RoBERTa*) using RoBERTa-base[2] model architecture from Huggingface [2]. We initialize the model with an off-the-shelf RoBERTa-base model based on the hypothesis that a conversational language model would be better able to generalize by leveraging existing language model properties learned by RoBERTa-base. The training is carried out on Masked Language Modelling task for 5 epochs on a V100 GPU with a learning rate, batch size and warmup steps of 5e-4, 8 and 100 respectively.

3.2. Fine-tune Task Specific Models

The fine-tuning approach for Silence-RoBERTa is as follows:

3.2.1. Data

For silence understanding tasks we sample 5.9K silent segments that are greater than 3 seconds from contact center conversations. For each of these segments we extract *L* words before the silence (left context) and *R* words after the silence (right con-

[1] We cannot release the datasets due to proprietary reasons.

[2] https://huggingface.co/roberta-base

Table 2: *Macro F1: Silence-Type Classification Task*

Approach	**Left Only**	**Left + Right**
Duration Based	–	49.14%
SVM	70.48%	67.69%
RoBERTa-base	81.25%	80.65%
ConvRoBERTa	83.19%	82.25%
Silence-RoBERTa	85.83%	84.37%

Table 3: *Macro F1: Causer Identification Task*

Approach	**Left Only**	**Left + Right**
Duration Based	–	44.50%
SVM	54.04%	62.15%
RoBERTa-base	63.26%	66.14%
ConvRoBERTa	72.85%	74.20%
Silence-RoBERTa	73.05%	76.03%

text) to create a unified window of left context, silence and right context. A group of three annotators is asked to annotate these windows with silence type labels - *Expected* or *Unexpected Silence* and silence causer labels - *Agent* or *Customer*. We observe an inter-annotator agreement [3] of 0.76 and 0.71 for the silence-type classification and causer identification tasks, respectively. The annotated dataset is then split into train, validation and test sets of 4.1k, 0.9k and 0.9k data points respectively.

3.2.2. Fine-Tuning Implementation Details

We use the pre-trained Silence-RoBERTa model obtained in Section 3.1.3 for task-specific fine-tuning. We choose two setups for feature extraction: 1) *Left Only* - Where only left context of silence is used, and 2) *Left + Right* - Where both left and right contexts are used. The choice of these setups is motivated from the differences between real-time and post-call settings. Real-time system is focused towards prompting agents with an alert for them to course correct at the onset of unexpected silence. In this case, only left context is available and hence we aim to study the performance of our system in *Left Only* setup. On the other hand, post-call analysis is done after the completion of call when entire transcript is available for analysis, hence providing the liberty to use both left and right contexts.

In both the setups, we encode contexts as per Section 3.1.2. It is to be noted that left and right contexts might contain additional silences apart from the one under consideration. For *Left Only* setup, we obtain <s> token representation and feed it to the classification layer. Whereas for *Left + Right* setup, we obtain independent <s> token representations for left and right contexts and concatenate them before feeding to the classification layer. The entire network is then trained with cross-entropy loss on a T4 GPU. We perform a hyper-parameter sweep over: learning rate $\in \{1e-5, 5e-5, 1e-4\}$, batch size $\in \{32, 64\}$, epochs $\in \{5, 10\}$ and weight decay $\in \{0.001, 0.01\}$ and choose the best setting based on Macro F1 on validation set.

4. Experiments and Results

We benchmark Silence-RoBERTa model against following baselines for the two tasks discussed in Section 2. For extracting left context, we find that *L=40* empirically performs best for *Left Only* setup. In *Left + Right* setup, we fix *L=40* and vary *R* from 5 to 20 tokens and find that *R=15* results in best performance for the tasks in Section 2.

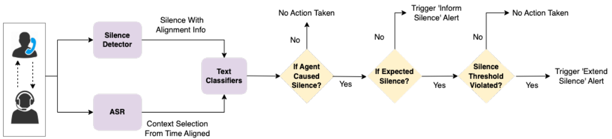
Silence Detector
Silence With Alignment Info
ASR
Context Selection From Time Aligned Transcript
Text Classifiers
If Agent Caused Silence?
No
No Action Taken
Yes
If Expected Silence?
No
Trigger 'Inform Silence' Alert
Yes
Silence Threshold Violated?
No
No Action Taken
Yes
Trigger 'Extend Silence' Alert

Figure 2: *Illustrative example of end-to-end pipeline for real-time monitoring of conversational silence*

1. **Duration based**: We choose this baseline since contact centers often focus on longer silences as it affects AHT and customer experience. Also, an exploratory analysis shows that *Unexpected* and *Customer* caused silences have lower average duration as compared to *Expected* and *Agent* caused silences. Hence, we label every silence greater than *K* seconds as *Expected Silence* caused by *Agent*, where *K* is chosen to maximize the Macro F1 on the validation set for both tasks.
2. **SVM**: We train a SVM [4] model with TFIDF features to report baseline with bag of words approach.
3. **RoBERTa-base**: We fine-tune RoBERTa-base model on silence understanding tasks to obtain out-of-box performance.
4. **ConvRoBERTa**: Since our data belongs to noisy ASR transcripts of conversations, we pre-train RoBERTa on in-domain data adopting similar pre-processing methodology as in Section 3.1.2 without encoding silence tokens. The obtained model is then fine-tuned on the two silence tasks.

Tables 2 and 3 list the results in above setups.

Silence-Type Classification: For silence-type classification task, we observe a higher F1 based on a bag-of-words SVM model as compared to duration based baseline, reinforcing our hypothesis that surrounding dialogue turns help in understanding conversational silences. Furthermore, fine-tuning RoBERTa-base model results in absolute improvement of 10-30% over duration based and SVM baselines in *Left Only* and *Left + Right* settings. Fine-tuning ConvRoBERTa model outperforms RoBERTa-base in both settings, signifying the impact of in-domain pre-training. Finally, we obtain 3-5% improvement over RoBERTa-base by fine-tuning Silence-RoBERTa model which further emphasizes the importance of using silence tokens in pre-training stage to capture the nuances of human conversations in contact centers.

Causer Identification: Similar to silence-type classification, SVM model outperforms duration based classifier on causer identification task, further exemplifying the usefulness of surrounding dialogue turns in understanding silences. We fine-tune RoBERTa-base, ConvRoBERTa and Silence-RoBERTa models on causer identification task and find a trend similar to silence-type classification in F1 score. Specifically, fine-tuning Silence-RoBERTa improves the F1 score by more than 9% as compared to RoBERTa-base model.

Real-Time Implications: Categorization of silence into expected vs unexpected depends on utterances before silence. Hence, for silence-type classification task, one can only utilize left context, allowing us to extend it to real-time setting. Results in Table 2 justify our hypothesis since the models in *Left Only* setup consistently outperform those in *Left + Right* setup. Conversely, adding right context generally improves the performance on causer identification task. Specifically, we observe *Left + Right* setup outperforms *Left Only* setup by an absolute margin of 3% using Silence-RoBERTa. However, with a slight trade-off in performance, it can be extended to provide feedback to agents in real-time where only left context is available.

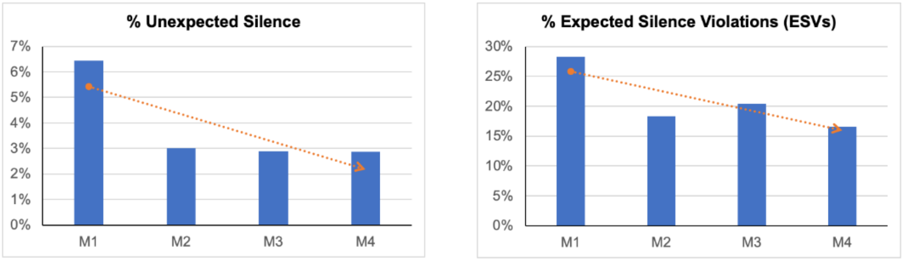

Figure 3: *Month-over-month comparison of KPIs associated with conversations silences. We observe a drop of 55.20% in % Unexpected Silences and 44.38% in %ESVs relative to the first month.*

5. Business Impact Assessment

In addition to the results in Section 4, we study the effectiveness and business impact of the proposed framework in a real-world contact center setup. Thus, we design a pilot study wherein we deploy the proposed silence-type classification and causer identification models with our internal CCaaS (Contact Center as a Service) platform. Refer to Figure 2 for schematic view of the platform. The system detects silences longer than 3 seconds and uses the models to infer silence-type and causer labels, which are then used to display alerts to agents as follows:

- **Scenario 1:** Causer is *Agent* and Silence-type is *Unexpected Silence*
  - We display an INFORM SILENCE alert (*"Inform the customer to expect a silence"*) to the agent
- **Scenario 2:** Causer is *Agent* and Silence-type is *Expected Silence*
  - We track the silence duration and if it exceeds a fixed threshold (e.g 60 seconds), we display an EXTEND SILENCE alert (*"Inform the customer if you need more time"*) to the agent
- **Scenario 3:** Causer is *Customer*
  - No action taken.

As a part of our study, we onboard a group of 10 contact center agents who are familiar with the interface of the platform. These agents are exposed to the above system that triggers real-time alerts at the onset of silences. In order to accurately assess impact of the proposed system, we ensure that agent experience with respect to user-interface of the CCaaS platform prior to and during the study remains identical, except for the inclusion of real-time alerts. We randomly assign incoming calls to these agents and monitor the following metrics on a monthly basis:

Percent Unexpected Silences: We define *Percent Unexpected Silences* as the proportion of total calls that consist of one or more instances of *Unexpected Silences*. We hypothesise that INFORM SILENCE alert provides real-time feedback to agents to proactively inform the customer to expect a silence. The results in Figure 3 justify our hypothesis wherein we observe a decreasing trend in the *Percent Unexpected Silences*, implying that exposure to these alerts drives better adherence to protocol by the agents while handling conversational silences.

Percent Expected Silence Violations (ESVs): We define *Percent ESVs* as the proportion of *Expected Silences* that required triggering an EXTEND SILENCE alert. We specifically call this as *Expected Silence "Violation"* as the agent has violated an acceptable silence duration (here, 60 seconds). Figure 3 shows a decreasing trend in *Percent ESVs* exemplifying the effectiveness of the alert in reducing proportion of longer silences which could potentially lead to poor customer experience.

6. Prior Work

Over the years, improving quality of service and customer satisfaction have been key focus areas for contact centers. Roy et al. [5] proposed real-time quality assurance system using statistical and rule-based NLP to enable agents' supervisors to monitor ongoing calls and take corrective actions. Ando et al. [6] proposed a joint modelling of turn-level and call-level estimation of customer satisfaction using LSTM-RNN. Additionally, Segura et al. [7] use CNNs on raw audio signals to learn features that help predict customer satisfaction in contact center calls.

Conversational silences are widely studied by researchers. While silence is defined as absence of speech, [8] argue that its presence often complements surrounding speech. In contact center domain, Chowdhury et al. [1] use dialogue turns around silence to study its functions towards information flow in a dyadic conversation. While there is significant work on understanding functions of silences, its impact on customer experience in contact centers space is relatively less studied.

Recent developments in language modeling using Transformer [9] based models like GPT [10], BERT [11], RoBERTa [12], etc. have led to significant improvement in performances on downstream NLP tasks. Significant efforts are being made to pre-train models using phoneme sequences that are robust to ASR errors and result in further improvements in downstream tasks on ASR transcribed texts [13, 14]. Kumar et al. [15] performs an investigative study by probing BERT based language models trained on spoken transcripts to understand its ability to learn multifarious properties in absence of speech cues.

7. Conclusion

In this paper, we put forth an effective framework for understanding conversational silences in contact centers using surrounding dialogue turns via two text classification tasks - *Silence-Type Classification* and *Causer Identification*. These would not only help contact centers surface silences that lead to poor customer experience but also help take strategic measures to coach agents. While language models generally discard silences in conversations, our study shows that encoding them helps in learning silence-aware representations for contact center conversations. Fine-tuning these silence-aware models not only leads to performance gains on our silence tasks but could also potentially benefit other downstream tasks. Furthermore, our methodology is applicable in understanding silences in both real-time as well as post-call analysis that not only helps improve customer experience but also boosts agent performance.

8. References

[1] S. A. Chowdhury, E. A. Stepanov, M. Danieli, and G. Riccardi, "Functions of silences towards information flow in spoken conversation," in *Proceedings of the Workshop on Speech-Centric Natural Language Processing, SCNLP@EMNLP 2017, Copenhagen, Denmark, September 7, 2017*, N. Ruiz and S. Bangalore, Eds. Association for Computational Linguistics, 2017, pp. 1–9. [Online]. Available: https://doi.org/10.18653/v1/w17-4601

[2] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, and J. Brew, "Huggingface's transformers: State-of-the-art natural language processing," *CoRR*, vol. abs/1910.03771, 2019. [Online]. Available: http://arxiv.org/abs/1910.03771

[3] J. Cohen, "Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit." *Psychological bulletin*, vol. 70, no. 4, p. 213, 1968.

[4] C. Cortes and V. Vapnik, "Support-vector networks," *Machine learning*, vol. 20, no. 3, pp. 273–297, 1995.

[5] S. Roy, R. Mariappan, S. Dandapat, S. Srivastava, S. Galhotra, and B. Peddamuthu, "Qa rt: A system for real-time holistic quality assurance for contact center dialogues," in *Thirtieth AAAI conference on artificial intelligence*, 2016.

[6] A. Ando, R. Masumura, H. Kamiyama, S. Kobashikawa, Y. Aono, and T. Toda, "Customer satisfaction estimation in contact center calls based on a hierarchical multi-task model," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 715–728, 2020.

[7] C. Segura, D. Balcells, M. Umbert, J. Arias, and J. Luque, "Automatic speech feature learning for continuous prediction of customer satisfaction in contact center phone calls," in *International Conference on Advances in Speech and Language Technologies for Iberian Languages*. Springer, 2016, pp. 255–265.

[8] A. Jaworski, *The power of silence: Social and pragmatic perspectives*. Sage Publications, 1992.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, Eds., 2017, pp. 5998–6008.

[10] A. Radford and K. Narasimhan, "Improving language understanding by generative pre-training," 2018.

[11] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, J. Burstein, C. Doran, and T. Solorio, Eds. Association for Computational Linguistics, 2019, pp. 4171–4186. [Online]. Available: https://doi.org/10.18653/v1/n19-1423

[12] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized BERT pretraining approach," *CoRR*, vol. abs/1907.11692, 2019. [Online]. Available: http://arxiv.org/abs/1907.11692

[13] M. N. Sundararaman, A. Kumar, and J. Vepa, "PhonemeBERT: Joint Language Modelling of Phoneme Sequence and ASR Transcript," in *Proc. Interspeech 2021*, 2021, pp. 3236–3240.

[14] Q. Chen, W. Wang, and Q. Zhang, "Pre-training for spoken language understanding with joint textual and phonetic representation learning," in *Interspeech 2021, 22nd Annual Conference of the International Speech Communication Association, Brno, Czechia, 30 August - 3 September 2021*, H. Hermansky, H. Cernocký, L. Burget, L. Lamel, O. Scharenborg, and P. Motlícek, Eds. ISCA, 2021, pp. 1244–1248. [Online]. Available: https://doi.org/10.21437/Interspeech.2021-234

[15] A. Kumar, M. N. Sundararaman, and J. Vepa, "What bert based language model learns in spoken transcripts: An empirical study," in *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, 2021, pp. 322–336.