



Into the Sound: Analysis of Technical Quality of Audio and Speech

Dayana Ribas¹, Juan Antonio Navarro¹, Fernando Macías², Luis Guillen Civera¹, José Javier Castejón¹, Fernando Barreiro-Lostres¹, Oihane Albizuri¹, Andrés Carrión¹, José María Vinacua¹, Luis Benavente¹, Martín Sagardía², José Luis Cortina², Juan Luis Moreno³, José Ángel de la Cruz³

¹BTS - Business Telecommunications Services

²Neovantas Consulting

³TCRM, Telefonica

dribas@bts.io, fmacias@neovantas.com

Abstract

This paper introduces Into the Sound (ITS), a toolkit developed for analyzing the technical quality of audio and speech in contact center interactions. Unlike traditional approaches that prioritize text transcriptions, ITS focuses on direct analysis of the audio signal, incorporating metrics related to audio quality, call flow, paralinguistics, and also transcription. By emphasizing audio characteristics, ITS provides a comprehensive assessment of both the technical and communicative dimensions of customer service interactions.

Index Terms: contact center, speech quality, call flow, paralinguistics

1. Introduction

Speech analytics refers to the systematic analysis of call center interactions through the examination of transcribed calls. Several companies¹ provide this service, enabling businesses to gain insights into customer sentiment, behavior, and emerging trends by analyzing key phrases, emotions, and topics discussed during conversations. These services often integrate with applications such as customer service platforms and business intelligence tools.

Although the primary data generated by call centers consists of audio from telephone conversations, speech analytics applications typically focus on processing the textual information derived from transcriptions, leaving much of the audio data underutilized. However, with advancements in artificial intelligence, particularly Natural Language Processing (NLP), the capabilities of speech analytics have significantly improved. These technologies enhance the automatic detection of patterns, identification of areas requiring improvement in customer service, and even the prediction of potential customer issues. Consequently, some companies have leveraged these advancements to improve decision-making processes and enhance overall customer experience.

This paper introduces the toolkit Into the Sound (ITS), developed as part of a project aimed at analyzing the technical quality of audio and speech within a contact center environment. The study was conducted over two months in 2024, in response to a request from the TCRM department at Telefonica². The department expressed interest in evaluating how the

¹Some prominent examples include: Verint Systems, Genesys, Avaya, among others.

²Spanish multinational telecommunications company: <https://www.telefonica.com>

technical quality of audio services affected the overall performance of contact center interactions. ITS was designed to analyze data directly from the audio signal, shifting the primary focus away from text transcriptions, which are typically the main source of information in traditional speech analytics applications. Instead, the transcription serves as a supplementary component. The toolkit integrates various metrics, including audio quality, call flow, paralinguistics³, and the transcription itself. This comprehensive analysis allows for a more nuanced understanding of both the technical and communicative aspects of the contact center's performance.

From this point forward, Section 2 comments on the applications of ITS toolkit within contact centers. Section 3 outlines the data material utilized in this project. Section 4 details the structure of the ITS, which consists of machine learning and signal processing techniques for analyzing audio files. Section 5 describes the system architecture, which is divided into a back-end responsible for implementing the methods discussed in the previous section, and a website for visualization purposes. This section also includes additional implementation details regarding the computation of metrics. Section 6 presents the results of the study using ITS toolkit to evaluate the impact of audio quality in the service of the contact center. Finally, Section 7 concludes the paper and provides guidelines for future work.

2. Applications of ITS toolkit

The ITS toolkit has a wide range of applications within contact centers, offering a comprehensive assessment that extends beyond simple text transcriptions. Some of its key applications include:

- 1. Identification of technical issues in contact center channels and systems:** ITS enables the differentiation of audio quality problems stemming from various sources, such as:
 - The communication channel (interference, signal loss).
 - Recording systems (audio compression, recording errors).
 - Background noise or interference in the agent's environment.
 - Agent equipment (issues with microphones or headsets).
- 2. Comparison between groups:** ITS facilitates the compar-

³Paralinguistic features refer to the aspects of spoken communication that go beyond the literal meaning of the words, focusing on how something is said rather than what is said. These features are crucial in conveying emotions, attitudes, and nuances in spoken interactions. Common paralinguistic features include pitch, speaking-rate, pauses, silence, etc

ison of quality metrics across different telecommunications service providers, contact centers, or even between individual agents. This helps to identify opportunities for improvement while ensuring that services meet the required standards at all levels.

3. **Monitoring indicators over time:** By analyzing metrics over several weeks or months, ITS helps track the evolution of service quality and detect patterns or trends that may point to areas needing improvement. This feature is particularly useful for evaluating the impact of operational or infrastructure changes.
4. **Alarm detection and report generation:** ITS automatically triggers alarms when certain thresholds are exceeded. This threshold will be defined by the customer and related to the reasonable value of metric. For instance, for the technical quality of the telephone calls, Signal to Noise Ratios (SNR) behind 5 dB or more than 20% of the call duration with saturation will trigger the alarm. Additionally, it produces customizable reports that offer detailed analyses by contact center, agent, recorder, or switch, making it easier to pinpoint specific issues in each category.
5. **Comprehensive audio quality evaluation:** Combining metrics of audio quality, call flow, and paralinguistic aspects, ITS provides a complete view of both technical performance and communicative effectiveness during interactions in the contact center. This analysis helps identify improvement areas in both agent performance and customer experience.
6. **Artifact analysis:** ITS can identify specific elements within the audio and parameterize them for tracking and analysis, such as distinct sounds, laughter, breathing, or remote work environments. During the pilot exercise, for instance, ITS enabled the comparison of audio quality variables in different work environments (office vs. remote work). Although this analysis was theoretical, it provided a proxy to identify the call's origin (agent's environment) and laid a foundation for evaluating the potential impact of remote work on service quality without the need for direct actions in contact centers.

3. Data

This project utilized a non-public dataset, referred to as MO-VISTARDB, which consists of telephonic audio recordings from customer service calls made to Movistar's well-known 1004 helpline. These recordings capture conversations between agents and customers in a realistic contact center scenario. The audio files are sampled at a rate of 8000 Hz, with 16-bit resolution, recorded in stereo, and saved in WAV format without additional encoding (signed-integer PCM format). The typical duration of these conversations is around 10 minutes, though some interactions are significantly shorter or consist of empty files, which were classified as part of a "trash set." Prior to recording, customers were informed that their calls would be recorded. All agents involved were adult native Spanish speakers, with a variety of accents from both Iberian and Latin American regions.

The study spanned two months, capturing audio from a total of 10,925 conversations, amounting to approximately 180,000 minutes of audio, which represents 1% of the full service volume. This data was collected from four contact centers, involving a total of 3,153 agents. The audio data was analyzed in four stages, with reports provided every two weeks, culminating in a final report summarizing the entire study.

4. Methodology

The core methodology of ITS is illustrated in Fig. 1. It comprises an ensemble of machine learning and signal processing techniques designed to facilitate the end-to-end processing of a set of audio files, with the objective of producing a report on the technical performance of the entire dataset.

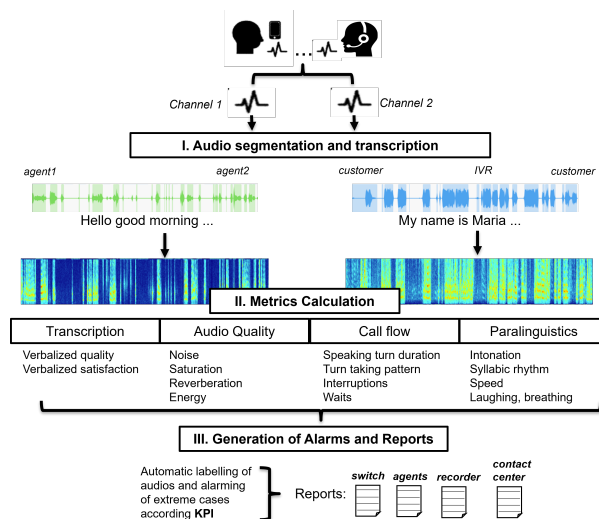


Figure 1: Schematic diagram of ITS.

The first stage involves pre-processing the audio files, which includes segmenting the audio and transcribing each channel of the telephone call. The segmentation tool is based on the Pyannote open-access toolkit for diarization [1], with further modifications to label segments containing IVR interactions and music. The transcription process utilizes the large model from OpenAI's Whisper open-access toolkit [2].

In the second stage, metrics are computed from both the audio and the corresponding text. These metrics are categorized into four thematic blocks: audio quality, call flow, paralinguistics, and transcription. The transcription-related metrics align with the current state-of-the-art technology commonly used by companies providing Speech Analytics services. They capture the verbal expressions of quality and satisfaction from the client or agent, as exemplified by utterances such as "I'm not at all satisfied with the management..." or "can't hear you well, the call is breaking up...".

In terms of audio-related metrics, the audio quality block identifies aspects of the call that indicate poor audio performance, such as saturation, reverberation, and audio artifacts. The call flow block analyzes the structure of the conversation and the interactions between speakers, measuring factors like response times, overlaps, and the duration of speaker turns. Finally, the paralinguistic block focuses on recognizing non-linguistic aspects of the speakers' communication, such as variations in pitch and syllabic rhythm.

The third stage involves combining all computed metrics to evaluate the overall performance of the service based on the processed data. This includes generating alarms when metrics exhibit extreme values, which are used to represent the technical quality Key Performance Indicator (KPI). To determine the alarms for each audio file, the system defines a set of indices linked to the KPI that provide an interpretation of the metrics' performance.

As illustrated in Fig. 2, the system defines several indices with typical values for this dataset. These include a general index (IPA), which triggers an alarm for an audio file if any of its metrics deviate beyond the mean \pm standard deviation of the dataset’s overall performance. Another index is the quality index (QI), which represents a weighted sum of the metrics within the audio quality block. Additionally, a satisfaction index (YES) is calculated, grouping and weighting the metrics associated with conversation performance.

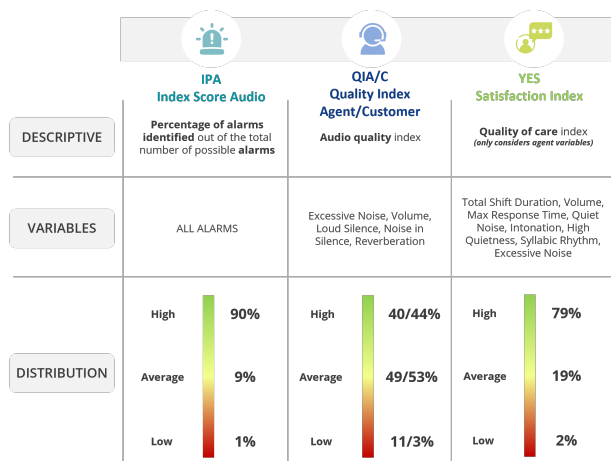


Figure 2: Indices employed for summarizing the performance of audio and text metrics according to KPI of the callcenter service.

Finally, all the information about the audio files compiled during the previous processing stages is organized and presented in reports, categorized by areas of interest: contact center, switch, agent, and other relevant categories specified by the contact center manager. This categorization is based on metadata that accompanies the audio files.

Note that this project primarily focused on evaluating the technical quality performance of the service, in line with the client’s specific interests at the time. However, the third stage of the reporting process can be customized to offer alternative interpretations of the audio and text metrics. This flexibility allows the final report to be tailored to the client’s evolving priorities. For example, in addition to assessing the technical quality of the calls, the ITS toolkit can be adapted to address broader operational business KPIs, such as sales performance, customer satisfaction, and productivity. This makes it possible to integrate the results into the business intelligence tools commonly used by call centers.

5. System and performance

This section aims to describe the system architecture, which is divided into a back-end that contains the code implementing the methods outlined in Section 4, and a front-end website designed for visualization purposes. Additionally, this section will include further implementation details regarding the overall computation process.

5.1. Backend

The backend containing the methods for processing audio and computing metrics is programmed in *Python*. The system is based on micro-services architecture, composed of a cluster

based on Kubernetes technology⁴ with different containers that operate in “session stateless” mode, which provide redundancy and scalability. Each of these containers stores the speech representation and machine learning methods employed for the audio processing. To provide the highest possible performance, especially in internal communications, the gRPC protocol⁵ was used, developing REST interfaces⁶ on top to allow external consumption. When executing the algorithms, GPU processing cards were used.

5.2. Visualization

The ITS features a web service designed to provide an user interface for information and visualization purposes, which consists of two primary components. The first component is a dashboard implemented using Apache Superset technology⁷, as illustrated in Fig. 3. This dashboard displays reports organized by contact center, switch, agent, and other relevant categories. The data storage management for the Superset visualization can be fully customized to match customer use cases. This is based on the Apache Pinot database⁸, providing efficient columnar storage and management for handling large volumes of data. The dashboard allows for easy customization by the client, enabling them to filter and select the data they wish to display through the options available in the left column.

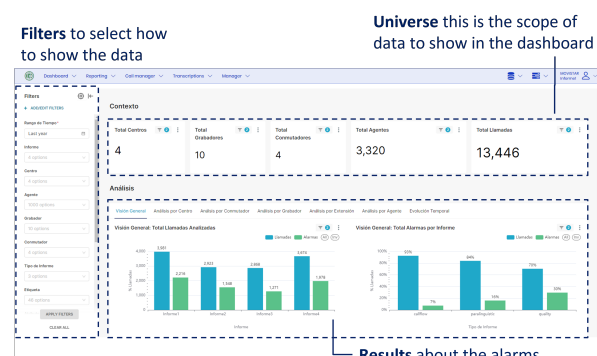


Figure 3: Dashboard of reports of ITS.

Additionally, Fig. 4 presents a sketch of the dashboard in its evolutionary mode, which allows users to delve deeper into all available levels and facilitates an over-time analysis of the service. Then, the audio manager, illustrated in Fig. 5, enables users to view detailed information about each audio file and relevant metrics. This dashboard also provides a toolkit for manually entering data, allowing users to annotate the audio based on quality and paralinguistic aspects. These web services are accessible at <https://uat.intothesound.ai/> and require proper authentication for client access.

5.3. Implementation details

5.3.1. Computational resources

From the point of view of computational cost, the use of resources in a shared, balanced and scalable way allows the investment in hardware to be minimized while a progressive

⁴<https://kubernetes.io/docs/reference/kubect/>

⁵<https://grpc.io/>

⁶<https://restfulapi.net/>

⁷<https://superset.apache.org/>

⁸<https://pinot.apache.org/>

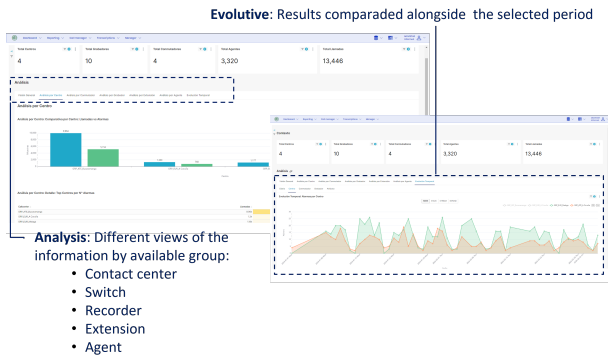


Figure 4: Audio manager of ITS with analysis over time.

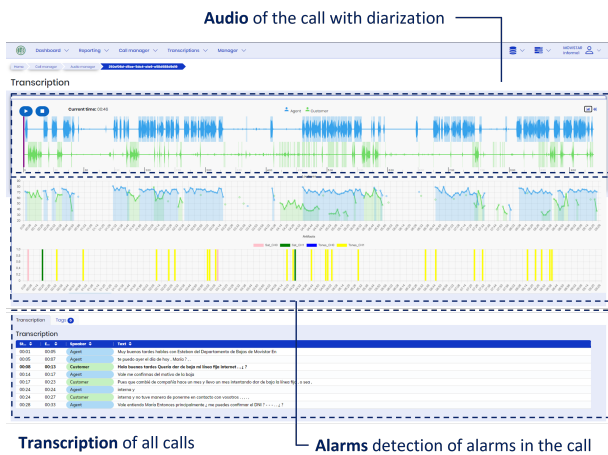


Figure 5: Audio manager of ITS.

adoption by clients allows the progressive acquisition or rental of this type of specialized hardware. This is especially relevant, if we take into account that the solution makes a very intensive use of hardware, especially the processing and memory capabilities of both the servers and dedicated GPU cards for computing artificial intelligence algorithms.

The analysis of audios required about 70 Gb of RAM per server, 32 CPUs (all) with Intel i9-13900 processor, 24 Gb of RAM in a GPU card: NVIDIA GeForce RTX 4090 plus all its computing capacity to perform audio processing. On the other hand, the adoption of industry standards, especially those based on open source, gives the project independence when choosing suppliers and developers, which results in a lower cost of ownership and its associated maintenance while empowering the research innovation and development of these kind of platforms.

5.3.2. Methodology for development

During the development and industrialization of ITS, we followed the Agile-Scrum methodology, as is typical for other projects within the company. This approach enabled us to ensure flexibility, iterative progress, and close collaboration with stakeholders. Specifically, during the two-month period when Telefonica was periodically sending audio packages, we found Agile-Scrum particularly useful for breaking down tasks into sprints. This allowed us to prioritize effectively and hold daily stand-ups to track progress. Each sprint concluded with a-

view and retrospective to assess results from the new audio package and incorporate them into the main pool of data. With each new audio package processed, we encountered fresh challenges, for instance, we discovered minor metric errors and unnecessary delays in implementations. This methodology enabled us to adapt to evolving requirements, deliver high-quality features incrementally, and maintain alignment with business goals throughout the development cycle.

6. Results and Discussion

This section presents results of a study conducted on the MO-VISTARDB dataset to evaluate the technical quality of audio files using the ITS toolkit. The findings aim to assess the impact of these quality metrics on the overall performance of the contact center service. However, it's important to note that the results are based on a dataset created specifically for this project, and may not accurately reflect the real-world performance of the company's services.

6.1. Alarms identified

Fig. 6 illustrates the total number of alarms generated across all analyzed calls. Notably, 52.9% of the audio files triggered an alarm for at least one metric, indicating that the IPA index was activated. Among these, 36.1% of the alarms originated from the agent channel, underscoring its sensitivity as this is the side where agents can actively work to improve service quality. In the lower section of the figure, the alarms are categorized by groups of metrics, revealing that audio quality is the most impacted dimension, with 34.8% of the audio files triggering alarms related to some quality-related metric.

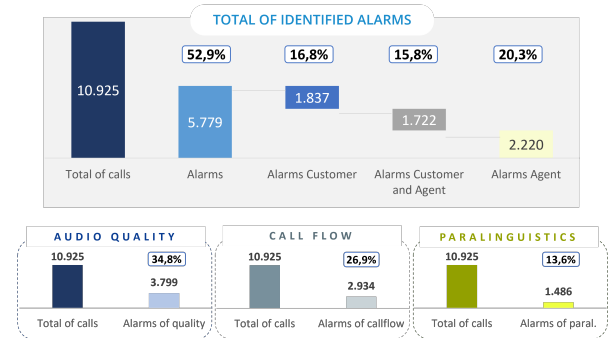


Figure 6: Total number of alarms.

6.2. Indices related to KPIs

Fig. 7 presents the performance of the indices related to KPIs over the course of several weeks.

At first glance, it is evident that the variation in the number of calls per week does not have a direct correlation with the behavior of the indices. The Quality Index for the agent channel consistently remains lower than that for the customer channel, reflecting the typical distortions present in the contact center environment. Nonetheless, this is a positive indicator, as it suggests there is room for improvement in service quality through targeted actions aimed at enhancing the working conditions for agents. Additionally, the graph reveals that the Satisfaction Index generally remains high. However, this graph does not provide clear evidence of a direct relationship with the Quality Index. Nevertheless, all indices tend to decrease as the

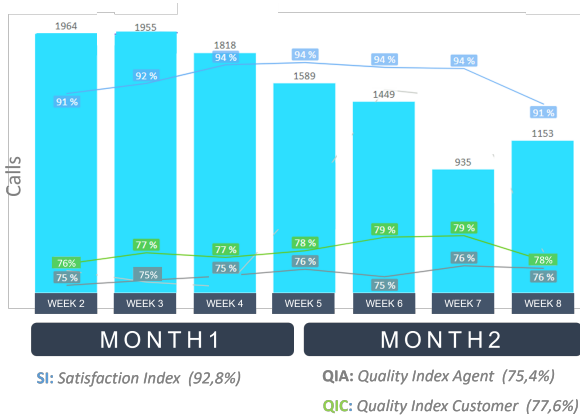


Figure 7: Results of indices related to KPI by week.

number of processed calls increases, which is expected, given that a higher call volume typically leads to more issues.

6.3. Comparative analysis by group

Fig. 8 presents a comparative analysis of the contact centers based on the indices related to KPIs. It is evident that Contact Center 2 exhibits the poorest performance regarding technical quality when compared to the other centers, as reflected in its lower Satisfaction Index relative to the others. This achievement aligns with the prior knowledge of the contact center managers regarding service performance, thereby validating our analysis.

	Calls	IPA Index	Satisfaction Index	Quality Index Agente
Contact Center 1	846 (7,7%)	97,6%	94,2%	79,8%
Contact Center 2	8.028 (73,5%)	96,5%	92,5%	73,0%
Contact Center 3	985 (9,0%)	97,6%	93,3%	83,0%
Contact Center 4	1.066 (9,8%)	97,3%	93,1%	82,8%

Figure 8: Comparative throughout contact centers.

Fig. 9 provides a comparative analysis of the recorders based on the indices related to KPIs. It is clear that Recorder 708355 demonstrates the poorest performance concerning both the IPA and Quality Index metrics.

	Calls	IPA	Quality Index (Agent and Customer)
708351	999 (9,1%)	96,5%	76,6%
708352	996 (9,1%)	96,8%	76,5%
708353	1.091 (10,0%)	96,8%	76,7%
708354	1.131 (10,4%)	97,0%	76,3%
708355	1.129 (10,3%)	96,6%	76,0%
708356	1.095 (10,0%)	96,4%	76,3%
708357	1.114 (10,2%)	97,0%	76,9%
708358	1.189 (10,9%)	96,8%	76,3%
708359	1.099 (10,1%)	96,8%	76,8%
708360	1.082 (9,9%)	96,6%	76,6%

Figure 9: Comparative throughout recorders.

Fig. 10 offers a comparative analysis of the switches based on the indices related to KPIs. Although the results are gener-

ally similar, Switch B exhibits a lower Quality Index compared to Switch A.

	Calls	IPA	Quality Index (Agent and Customer)
Switch A	2.976 (50,1%)	96,6%	76,1%
Switch B	2.986 (49,9%)	96,6%	75,9%

Figure 10: Comparative throughout switches.

Therefore, the comparative analysis of relevant categories in this study (contact centers, switches, recorders) highlighted by performance metrics, enables the identification of specific areas requiring targeted interventions to improve service quality, particularly in cases of lower Quality Indices. These insights are crucial for guiding strategic initiatives to optimize operations and, ultimately, enhance customer satisfaction across the service ecosystem.

6.4. Discussion of results

Previous results from the experience of using ITS toolkit with MOVISTARDB have demonstrated its capability to provide a comprehensive evaluation of service quality in contact centers, distinguishing between issues with different origins and impacts on the service. This exercise provided Telefonica with valuable insights into the potential of the tool and the specific data that can be leveraged to enhance its service quality. Some of the key points identified during the exercise were:

- Recorder and switch issues:** ITS detected significant issues with one of the recorders, validating an incident previously identified by the service team. Another recorder was also found to have atypical quality values, which were reported to the provider, who then proceeded with its review and adjustment.
- Quality differences between contact centers:** A comparative analysis between contact centers revealed problems in two of the centers evaluated. In the first, the issues were related to background noise, while in the second, problems were identified with the equipment used by agents. These findings corroborated the results of previous manual audits, facilitating the follow-up on corrective actions.
- Impact of quality on customer dissatisfaction:** The ITS analysis made it possible to measure how quality issues affect customer satisfaction. It was found that calls with alarms related to quality—such as noise, call flow issues like overlapping or excessive silence, and paralinguistic aspects such as syllabic rhythm—exhibited higher levels of dissatisfaction. This allows for the definition of corrective actions to improve the service.

7. Conclusions and Future

In conclusion, the ITS toolkit, based on Open Software and cutting edge artificial intelligence technologies, offers significant value for contact centers by providing comprehensive insights derived from audio information that extend beyond mere transcription techniques. By employing advanced analytics on audio quality, call flow, and paralinguistic features, the toolkit enables a nuanced understanding of interactions that informs both operational improvements and strategic decision-making.

This holistic approach not only enhances the evaluation of service performance but also supports targeted interventions aimed at improving customer satisfaction and overall service quality. The integration of these audio-driven insights positions the ITS toolkit as a useful resource for optimizing contact center operations.

About the results of the study in MOVISTARDB, we conclude that the comparative analysis of performance across switches, recorders, and contact centers reveals distinct variances in technical quality as indicated by the KPI indices. While the overall results show comparable performance among the evaluated entities, specific outliers, such as Recorder 708355 and Switch B, highlight areas that require attention and improvement. Notably, the performance metrics suggest that certain contact centers may benefit from targeted interventions to enhance service quality, particularly in cases where lower Quality Indices are observed. These insights can guide strategic initiatives aimed at optimizing operations and ultimately improving customer satisfaction across the entire service ecosystem. We would like to emphasize that this study was conducted using a dataset specifically created for the project purposes, therefore, the results obtained do not necessarily reflect the actual performance of the company's services.

In the future, ITS toolkit can also be applied to operational business KPIs commonly used in call centers to help in the decision making process, such as sales, customer satisfaction, and productivity. The integration of ITS to the business intelligence tool of the contact center enables real-time monitoring and analysis of key performance metrics, allowing businesses to gain deeper insights into their operational efficiency and customer interactions. By leveraging the data from the ITS toolkit, businesses enhance their accuracy acquiring information beyond the transcription, from the hole audio.

On the other hand, we plan to approach to the developing of robust security and privacy protocols for audio treatment within ITS. Specially for the contact center scenario, since there is handled a large volume of personal data through recorded conversations, therefore it is crucial to protect sensitive customer information and ensure compliance with data protection regulations.

8. Acknowledgements

The authors would like to acknowledge the Vivolab Research Group at the University of Zaragoza for their contributions during the initial design and implementation stages of the algorithms underpinning the ITS toolkit methodology. Their ongoing support and assistance with any related issues are greatly appreciated.

9. References

- [1] H. Bredin, "pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe," in *Proc. INTERSPEECH 2023*, 2023.
- [2] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022. [Online]. Available: <https://arxiv.org/abs/2212.04356>