WHITE PAPER

Customer Service Call Center
Virtual Voice Assistance

AI-Powered Next-Generation Contact Centers

Contact Center Transformation Enabled by Intel® Xeon® Scalable Processors

Table of Contents

Executive Summary 1

Contact Center Transformation . . . 1

Profits Are a Phone Call Away . . 1

AI Driving Contact Center Transformation 2

Business Impact of AI-Enabled Contact Centers 2

Requirements for Scalable Conversational AI Deployments . . 3

Real-Time Performance 3

Flexible Deployment Options . . 4

Dynamic Load Scaling 4

A Vibrant Ecosystem of Solution Providers 5

Optimizing the ASR Algorithm Reduces Processing Time 5

Conclusion 5

Executive Summary

The best experiences by customers influence their next spending opportunity. A primary channel for enterprises to influence customer experience is their interactions with a company's contact center. Conversational solutions, such as virtual voice assistant and speech analytics powered by artificial intelligence (AI), are beginning to transform modern contact centers and boost customer experiences.

But successful deployments of conversational AI solutions require platforms that are both performant and scalable to handle a massive number of concurrent calls quickly and efficiently.

This white paper looks at the transformation under way in AI-powered contact centers and the platform requirements for successful implementations. One such implementation shows how Scalable processors help accelerate conversational solutions from AI Builders member, Gnani.ai.

Contact Center Transformation

"Please wait for the next available agent. All calls are handled in the order they are received."

Long wait times to reach a contact center operator or an inefficiently handled service call can result in frustrated customers.

Profits Are a Phone Call Away

Callers need answers before they purchase, or they want solutions after they purchase. Sometimes it's a simple question that needs to be answered, but it results in a long thread of button pushing, voice responding, and ultimately landing in a queue to wait for a person. Their experience through this contact process—including the final person-to-person exchange—will drive their eventual impetus for first and repeat purchases. According to the Harvard Business Review, the best past experiences result in 140 percent more spending than bad past experiences.[1]

Artificial Intelligence (AI) Driving Contact Center Transformation

Contact centers are expensive to run and maintain. Essential as they are, operators are sensitive to the center's total cost of ownership (TCO). Thus, companies are exploring novel technology solutions to vastly improve the customer experience without driving up TCO. Algorithms that enable digital assistants like Amazon Alexa and Google Assistant are finding their way into the software that drives customer contact centers. Artificial intelligence (AI)-powered virtual voice assistants and speech analytics are transforming how contact centers will operate (Figure 1). These solutions enable 24/7 quality customer service, while also assisting agents and contact center operators to boost productivity.

Business Impact of AI-Enabled Contact Centers

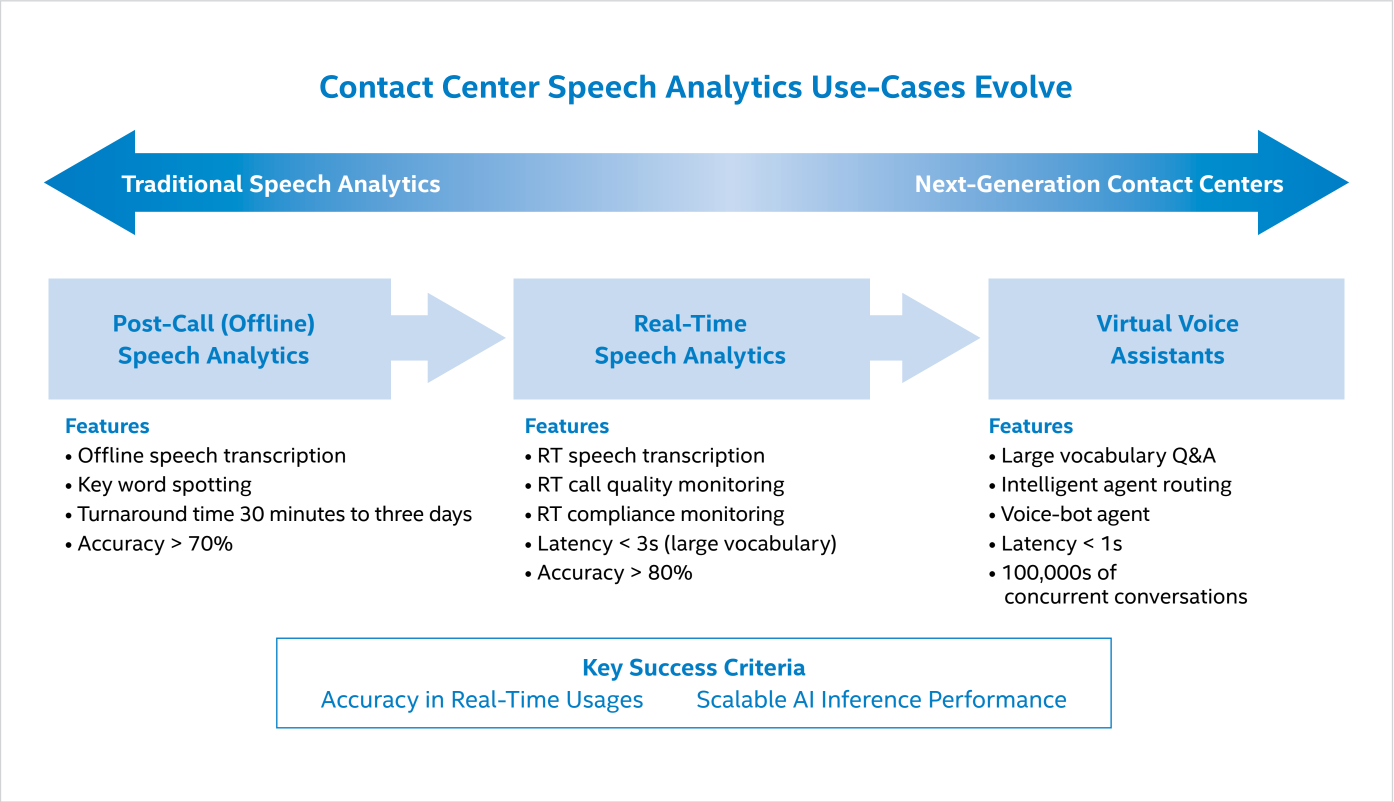
Call centers generate about 20 million hours of call recordings on a daily basis.[2] Traditionally, speech analytics has been applied offline on these large batches of recorded audio to gain insight into customer requests, problems, and interactions. The turnaround times for these analytics run from 30 minutes to three days.[3] The insights then drive the customer interaction scripts for the contact center—whether a human or programmed assistant bot provides the response.

Next-generation real-time speech analytics with very low latency and high response accuracy can allow businesses to quickly respond to customer concerns and questions.

AI-enabled virtual voice assistants integrate real-time analytics with a voice-bot agent to interact with callers, delivering an intelligent, fully automated experience. With such capability in the industry, Gartner estimates that by 2020, 25 percent of customer service and support operations will integrate AI-based virtual voice assistant technology across engagement channels.[4] This is expected to create a USD 7 to 20 billion industry, providing AI-powered solutions for customer service.[5]

> *"In present systems, customer service agents can spend a lot of time validating customers and understanding their queries. With voice bots, customer validation and even queries categorization for complex issues can be handled easily. This makes the whole process more efficient and reduces the wait times for customers, helping to deliver an ideal customer experience."*
>
> *– Ganesh Gopalan, CEO, Gnani.ai*

Compared to live operator calls, interactions handled by a virtual voice assistant can help reduce costs 75 to 86 percent (for five to 15-minute customer calls).[6] Virtual voice assistants also offer the benefit of handling calls around the clock in a geography-agnostic manner. Thus, AI-enabled call center solutions can significantly improve contact center cost and efficiency, while also ensuring fast and high quality resolutions.


Contact Center Speech Analytics Use-Cases Evolve
Traditional Speech Analytics
Next-Generation Contact Centers
Post-Call (Offline) Speech Analytics
Real-Time Speech Analytics
Virtual Voice Assistants

Features
- Offline speech transcription
- Key word spotting
- Turnaround time 30 minutes to three days
- Accuracy > 70%

Features
- RT speech transcription
- RT call quality monitoring
- RT compliance monitoring
- Latency < 3s (large vocabulary)
- Accuracy > 80%

Features
- Large vocabulary Q&A
- Intelligent agent routing
- Voice-bot agent
- Latency < 1s
- 100,000s of concurrent conversations

Key Success Criteria
Accuracy in Real-Time Usages
Scalable AI Inference Performance

Figure 1. Transforming future contact centers.

Requirements for Scalable Conversational AI Deployments

Fast, accurate inferencing, real-time analytics, and a large number of voice-bot instances to handle massive call volumes demand the highest level of performance, accuracy, and scalability in the contact center solution platform. Intel offers a portfolio of AI solutions that combine Intel Xeon Scalable processors, optimized Intel libraries, and software tools to address the needs of intelligent contact centers.

Real-Time Performance

Virtual voice assistant architecture is a complex integration of technologies (Figure 2). An end-to-end deployment of a solution brings together a variety of components that include algorithms with deep neural networks (DNNs), traditional speech processing, graph processing, enterprise software integration, and other functionality, with servers, storage, and networking. The computational requirements for minimizing latency throughout the pipeline are not trivial. Yet, using a technology to accelerate only one aspect of the overall solution, such as the DNN kernel using GPUs, does not fully address what the contact center operator and customers will care about—overall throughput.

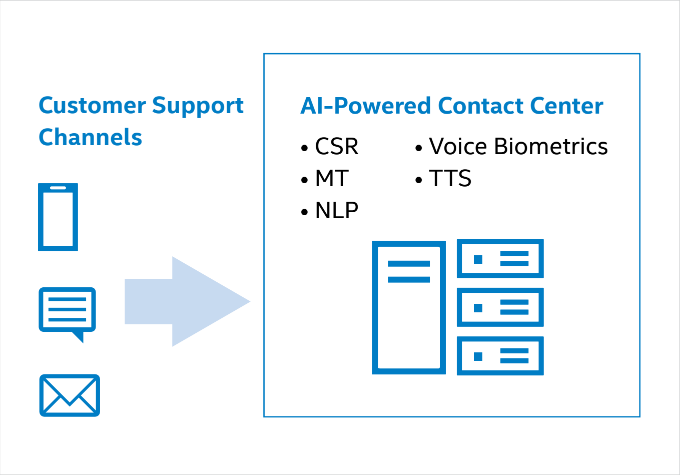


Figure 2. An AI-enabled contact center is a complex integration of technologies.

Processors Designed to Accelerate Conversational AI

The Intel Xeon Scalable processor family is designed for the most demanding workloads, delivering high-performance inferencing and scalability benefits to AI-enabled solutions, such as speech analytics. Additionally, hardware-enhanced technologies integrated in the processors enable scalable speech analytics solution deployments both in the cloud and on premise. Examples of these technologies include:

- DLBoost instruction[7] to accelerate inferencing.
- Intel Advanced Vector Extensions 512 (Intel AVX5212) for floating point math.
- Enabling Intel Optane™ DC persistent memory[7] for large memory pools or hierarchical memory.
- Intel Virtualization Technology (Intel VT) for Improved virtualization performance.

Software Optimized for AI

Conversational AI algorithms use diverse frameworks such as Kaldi toolkit, Pytorch, Caffe, TensorFlow, and others. The Intel AI portfolio includes several libraries, optimized software frameworks, and tools that significantly accelerate functions found in these frameworks and the programming languages often used in AI. Optimized libraries and tools include the following:

- Intel Math Kernel Library (Intel® MKL) with Basic Linear Algebra Subprograms (BLAS) optimizations
- Intel Math Kernel Library for Deep Neural Networks (Intel MKL-DNN) for neural network primitives
- Optimized memory manager libraries, like jemalloc or tbbmalloc
- Thread manager tools, like OpenMP and Intel Threading Building Blocks (Intel TBB)
- C and C++ compilers

The Intel AI software portfolio provides performance boosts across implementations for AI, including speech analytics algorithms, such as Automatic Speech Recognition (ASR), Natural Language Processing (NLP), Machine Translation (MT), Text-to-Speech (TTS), and Voice Biometrics. Such improvements help deliver real-time performance of the entire Conversational AI pipeline.

Flexible Deployment Options

Contact Centers have strict data privacy compliance requirements. Depending on the nature of data, these requirements regulate where (physical location) customer data is processed, stored, and purged. Virtual voice assistant applications must seamlessly execute across a range of infrastructure configurations, including on-prem enterprise data center, captive data center, or cloud data center, irrespective of the details in platform components. The platform architecture should support performance, power, and form factor requirements of the various deployment scenarios.

Intel Xeon Scalable processors are built on a common micro-architecture and instruction set architecture (ISA) over the entire range of product SKUs. Software developed for the ISA run consistently across the various models, enabling deployment of virtual voice assistant software from edge to cloud and on-prem without additional code changes. The hardware compatibility means Intel software libraries and tools used to optimize speech analytics workloads can be deployed flexibly across deployment scenarios (Figure 3).

Performance gains are available to the workload in a hardware-agnostic manner, greatly reducing development and validation efforts by software vendors.

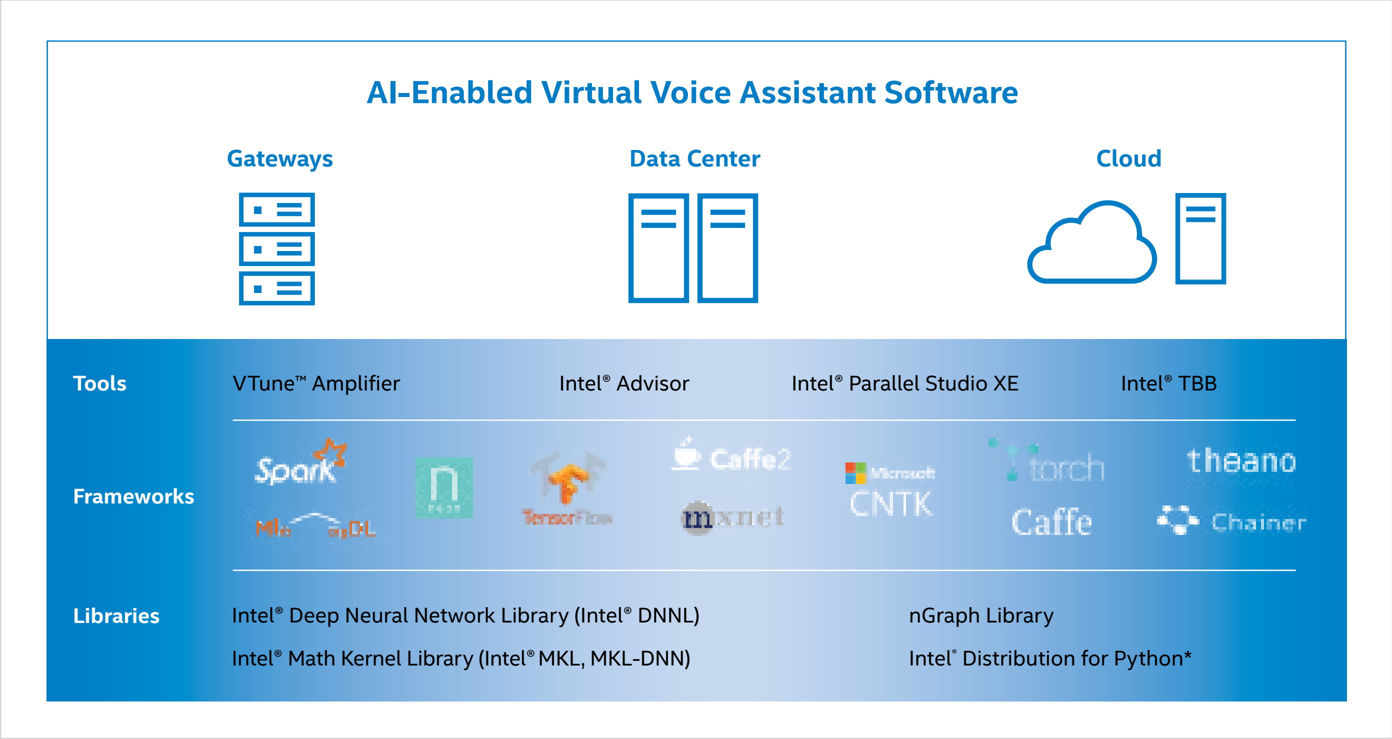
Dynamic Load Scaling

Contact centers experience variations in call volumes based on time of day, product launches, festivals, sales, and more. Calls can fluctuate from a few thousand to tens or hundreds of thousands, and back down in a short time. Quickly responding to the computational needs in a virtual voice assistant requires an agile and scalable infrastructure that will not shock the cost-sensitive contact center industry, which values low TCO.

High utilization efficiency of server resources as voice traffic and responses scale up and down can greatly impact TCO. Optimized software helps make more efficient use of the processor and platform pipeline, resulting in optimal hardware infrastructure needs as demand increases.

Intel offers tools for optimizing applications, such as VTune™ Amplifier, Intel Advisor, and Intel Parallel Studio XE. These help developers evaluate utilization of hardware resources and offer recommendations on boosting utilization. For applications based on Python, Intel offers Intel Distribution for Python to help accelerate codes on Intel processors. Thread manager tools, like OpenMP and Intel TBB, help improve software parallelism using multi-thread and multi-instance configurations. Implementing code optimizations based on these tools can boost hardware utilization significantly when the contact center is handling millions of concurrent voice channels globally.

These tools are available to developers, including solution provider partners, to build efficient and cost-effective Conversational AI systems.


AI-Enabled Virtual Voice Assistant Software

Gateways

Data Center

Cloud

Tools: VTune™ Amplifier; Intel® Advisor; Intel® Parallel Studio XE; Intel® TBB

Frameworks

Libraries: Intel® Deep Neural Network Library (Intel® DNNL); nGraph Library; Intel® Math Kernel Library (Intel® MKL, MKL-DNN); Intel® Distribution for Python*

Figure 3. Intel® Xeon® Scalable processors enable virtual voice assistant software to run consistently edge to cloud.

A Vibrant Ecosystem of Solution Providers

Customer service calls today handle conversations across hundreds of languages. Deployments tend to be very domain specific, designed to address unique product names, acronyms, and other characteristics and issues. A vibrant ecosystem of solution providers are servicing demand for Conversational AI across enterprise contact centers.

Gnani.ai (https://gnani.ai) develops domain-specific voice assistants to automate enterprise customer service. In addition to Gnani.ai's ASR and NLP engines, Gnani.ai adds vertical domain-specific and customer-specific intelligence to their platform to enable "automated voice bots" that can quickly resolve customer queries.

"The technical challenge that lies in voice bot adoption is the ability of ASR engines and NLP modules to recognize multiple languages, dialects, and accent variations of speakers and understanding the nuances and linguistic meanings of the domain. Gnani.ai technologies help efficiently address these challenges with a continuous learning layer to understand and learn the domain-specific intelligence."

– Ganesh Gopalan, CEO, Gnani.ai

Optimizing the ASR Algorithm Reduces Processing Time

Gnani.ai worked closely with Intel engineers to realize performance benefits for their solutions running on Intel Xeon Scalable processors. Automated Speech Recognition is the most compute-heavy part of the voice assistant pipeline. Improvements to ASR latency have a critical effect on the overall speech pipeline latency and, thus, was the focus of optimizations.

By utilizing Intel software libraries, compiler, and performance and parallelization tools, Gnani.ai achieved a notable reduction in processing time for the ASR algorithm This reduction allows their application to handle more streams in parallel on a given Intel-based platform.

Another key component of speech solutions is the language decoder. Intel provides optimizations for decoders, like CTC beam search and Viterbi algorithm-based decoders (using WFST to combine HMM information). Intel also offers its Intel Integrated Performance Primitives (Intel IPP) with optimizations to boost performance across various domains, including signal processing.

Gnani.ai also evaluated algorithm performance across different CPU SKUs, ranging from 8th Generation Intel Core processors to Intel Xeon Scalable processors. Tests showed similar performance boost across CPU SKUs before and after optimizations with no further software changes This allows ASR deployment across a range of infrastructure configurations, including on-prem enterprise data center, captive data center, or cloud data center.

Conclusion

Customer experiences with company service and support representatives impact how customers and potential customers make buying decisions. Supporting these voice interactions is necessary, but expensive, and enterprises are seeking automated, AI-enabled solutions that meet their sensitive TCO requirements. The new wave of AI-powered Virtual Voice Assistants is beginning to transform contact centers. Companies like Gnani.ai are advancing ASR and NLP technologies for virtual voice assistants and voice bots.

To succeed, a solution's entire processing pipeline needs to be fast, capable of efficiently handling multiple processes simultaneously. The platform must also be able to dynamically scale to support a massive number of concurrent voice channels when necessary. To deliver on the promise of AI-enabled speech analytics, it is imperative that solution providers deploy their software on platforms that provide scalable performance from the edge to the cloud or on-prem data center.

Using Intel Xeon Scalable processors and Intel optimizations, Gnani.ai was able to reduce processing time of their ASR algorithm, which allows more concurrent voice streams to be handled on the same platform. Intel Xeon Scalable processors offer unique performance benefits for this class of workload, delivering high performance and potentially lower TCO at peak concurrent usages.

Intel and Gnani.ai continue to optimize performance for speech analytics algorithms with the objective of accelerating the transition towards intelligent voice bot services and virtual voice assistants.

INTEL®
AI BUILDERS
MEMBER

Gnanai.ai is a member of the Intel AI Builders Program, an ecosystem of industry-leading independent software vendors (ISVs), system integrators (SIs), original equipment manufacturers (OEMs), and enterprise end users, which have a shared mission to accelerate the adoption of artificial intelligence across Intel platforms.

1. https://hbr.org/2014/08/the-value-of-customer-experience-quantified
2. Based on Intel analysis: www.vhtcx.com (5 hours on call per day) https://www.customerserv.com/blog/how-big-call-center-industry (approximately 4 Million workers with 20 million hours per day).
3. Intel analysis from customer inputs.
4. https://www.gartner.com/en/newsroom/press-releases/2018-02-19-gartner-says-25-percent-of-customer-service-operations-will-use-virtual-customer-assistants-by-2020
5. https://www.statista.com/statistics/589068/worldwide-virtual-digital-assistants-enterprise-market/
https://www.tractica.com/newsroom/press-releases/enterprise-virtual-digital-assistant-users-to-surpass-1-billion-by-2025/
https://www.marketsandmarkets.com/PressReleases/interactive-voice-response.asp
6. http://blog.genesys.com/enable-speech-recognition-in-your-contact-center/
7. Part of the 2nd Generation Intel® Xeon® Scalable processors.

0120/BA/HBD/PDF ♻ Please Recycle 341429-001US