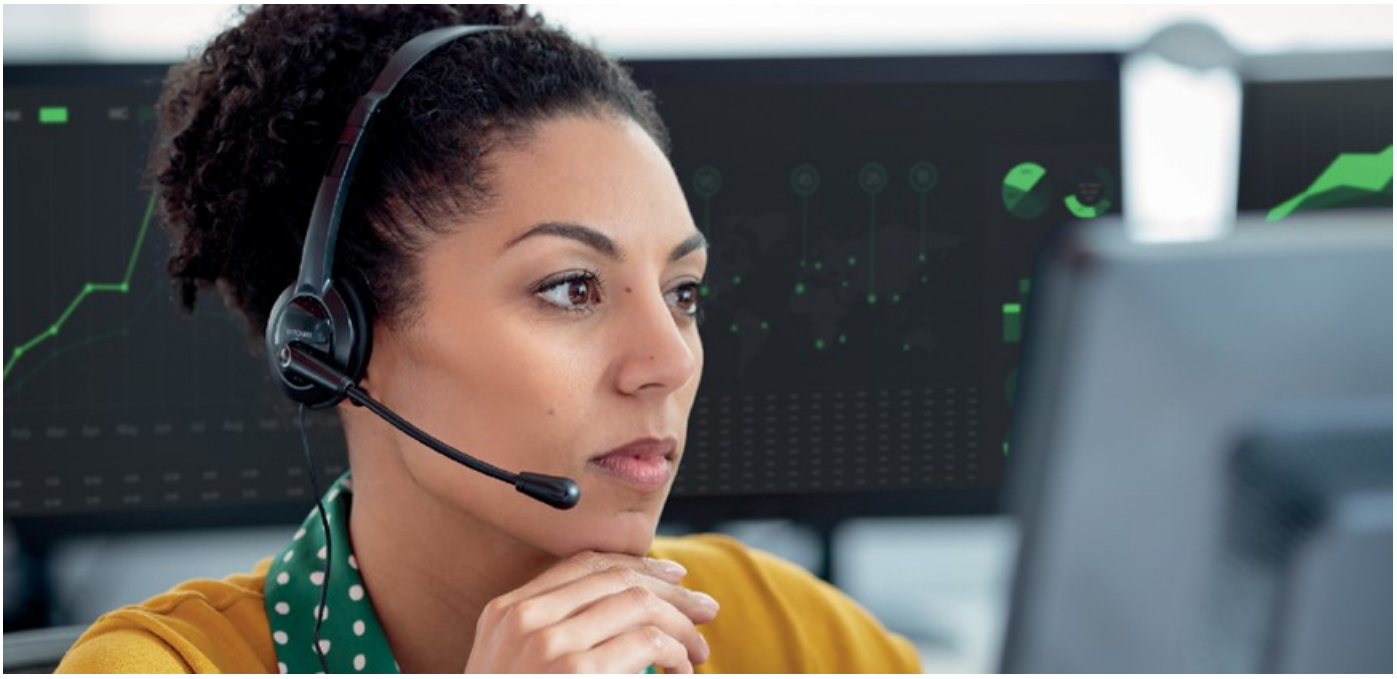
Hewlett Packard Enterprise

Technical white paper

SPEECH AND NATURAL LANGUAGE PROCESSING SOLUTION

Building an AI solution from edge to core



CONTENTS

1. INTRODUCTION

Artificial intelligence (AI)-powered natural language processing (NLP) represents a new era in human-computer interactions. Speaking to your computer, smartphone, or smartwatch is becoming increasingly commonplace—perhaps to the point one day, we may relegate the computer keyboard to a second-class peripheral like the once-ubiquitous CD drive.

Yet processing such a complex and deeply rooted human phenomenon as a language requires both strong computational power and intelligent NLP technologies, which makes deciding on an NLP solution for your organization an important one.

To begin with, which NLP solution best fits your organization's needs? Which hardware systems and accelerators best complement NLP technology? Which independent software vendors (ISVs) are most appropriate for your use case and geography?

Hewlett Packard Enterprise delivers outstanding expertise, technology, and AI/machine learning (ML)/deep learning (DL) components along with a global ecosystem of partners to help you unlock actionable insights and value.

This white paper covers both speech-to-text (STT) technologies, as well as the larger framework for NLP solutions. It also describes the infrastructure for the speech and natural language processing solution from HPE. The systems detailed here span from edge to core and edge to cloud as detailed in both reference configurations (Section 6) and representative industry use cases (Section 4). The framework is based on HPE hardware and software, which operates within the context of third-party ISVs and the HPE Pointnext Services organization.

2. NLP AND STT AT A GLANCE

Before analyzing the NLP marketplace in detail, a few basic terms should first be defined.

Natural language processing is the umbrella term that covers the entire field of AI-language analysis and generation that will be discussed in this paper.

The two main subfields of NLP are natural language understanding (NLU) and natural language generation (NLG)—each representing a linguistic interface between AI and the human subject. NLU travels one direction between AI and humans; NLG travels the other. With NLU, AI systems attempt to understand and process language generated by humans. With NLG, AI systems generate natural-sounding or seeming language for communication with listeners/readers.

If the language is spoken and not written/typed, then an additional layer is needed in the system. STT translates a human's spoken words into text on the screen. Crucially, there is no attempt at comprehension or meaning in the STT layer. It is rather acting as a stenographer, putting into text words being spoken. Of course, text to speech is the opposite, with the computer reading a text passage aloud.

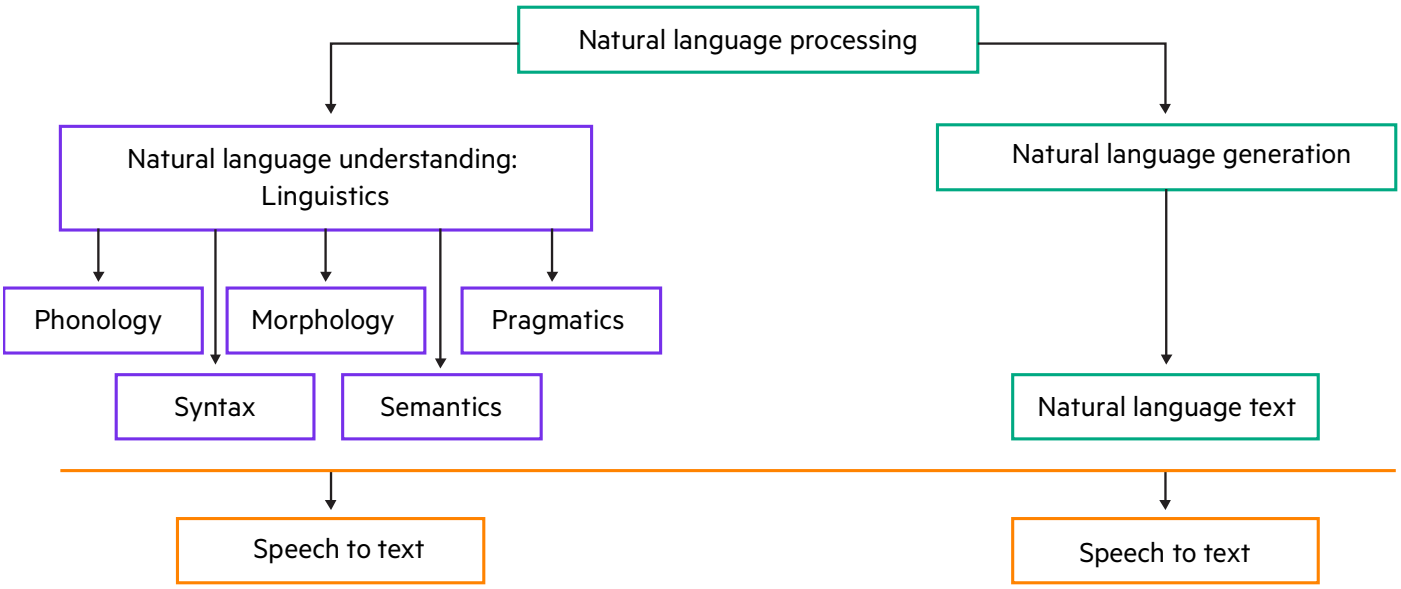


FIGURE 1. Data and analytics

2.1 Market

According to the analyst firm Tractica, global software, hardware, and services spending on NLP will reach $43.3 billion by 2025 enabling 44 discrete use cases across 17 industries.[1]

Some of the biggest drivers for STT are chatbots and voice recognition modules, which further enable digital assistants such as Siri, Alexa, and so on. Many companies, regardless of industry, are working to make their business voice-driven and -enabled.

Tractica's comprehensive market analysis explores some 43 NLP use cases in a range of industries, including:

TABLE 1. NLP use cases

FSI	Legal	Education	Medical/Healthcare	Customer service
E-commerce/sales virtual digital assistants	Patient research and analysis	Spoken fluency evaluation	Patient data processing	Virtual digital assistants
Tax filing and processing	Legal document review/research	Automated test grading	Medical treatment recommendation	Sentiment analysis
Financial search engines	Contract analysis	Education for autistic and speech-deficient students	Mining, processing, analyzing clinical notes	Text-based automated bots
Real-time competitive intelligence		Foreign language tutoring	Hospital patient management system	Social media bots

2.2 Challenges

Key challenges in NLP and STT involve technical, as well as legal considerations.

To transcribe, generate, or process language involves achieving a level of accuracy across seven levels of meaning and communication:

- **Phonological:** The sound of individual words
- **Morphological:** The units of each word (roots, prefixes, suffixes, and more)
- **Lexical:** The individual word
- **Syntactic:** The grammatical structure of a sentence
- **Semantic:** The possible meanings of a sentence
- **Discourse:** Meanings accrued across sentences
- **Pragmatic:** The entire sense of a passage, incorporating context, situation, intention, and more

As an example of these challenges, consider an STT application. First, the STT application must discern speech accurately from all background noise in a sound source. It must then extract individual words, although the speaker may unintentionally group some of them (for example, dancing and smile vs. dance, sing, and smile). There are, of course, many homophones for the spoken words in any given piece of text (red vs. read or they're vs. there vs. their, and so on). Slight differences in pronunciation can also denote marked departures in meaning (for example, talking in the library is always loud vs. talking in the library is always allowed). Then there are accents, dialects, individual speech patterns, and impediments, as well as differences in a speaker's age and emotional state that can also greatly affect the challenge an STT application faces in transcribing speech accurately.

STT is neither simple nor trivial. Yet ML, driven by the advent of accelerators, has enabled training of STT systems with increasing sophistication and accuracy. Depending on the amount of data and the size of the model, those accelerated systems can be powered by either GPUs or very powerful CPUs.

Unlike financial and banking transactions, STT and NLP communications do not typically include high levels of encryption, partly because of the high processing demands on STT and NLP systems at the network's edge. Most voice solutions today rely on the cloud to record and store voice data.

If the cloud provider acts as the conduit for all information to and from the consumer, which could include sensitive financial and health information, security risks should be considered. As most APIs for voice recognition are cloud-based, businesses building solutions in heavily regulated industries such as finance and healthcare should be proactive in their privacy and security practices. More and more

[1] tractica.com/newsroom/press-releases/natural-language-processing-is-a-key-engine-of-ai-market-growth-enabling-44-discrete-use-cases-across-17-industries/

industries must comply with stricter privacy regulations now that the EU's General Data Protection Regulation (GDPR) has been implemented.

Alternatively, HPE provides a scalable, secure speech recognition capability with cloud economics at the edge or on-premises for highly regulated and privacy-sensitive environments (see Section 3 for further information).

3. BUILDING AN NLP SOLUTION

HPE and Intelligent Voice pave the way for AI-powered NLP and STT solutions. Building an NLP solution begins at the level of the methodology used for text mining,[2] which is at the core of NLP system training. Methodologies are important because the size and scope of the text data sets will be beyond the capacities of individuals to manually intervene, audit, or process in any substantive way. Therefore, a standardized text mining methodology ensures the greatest reliability and automation for knowledge extraction from a given set of text resources. The data mining community has adopted cross-industry standard process for data mining (CRISP-DM) since 1999. Text mining is still a relatively new field and does not have similar international agreements in place. However, a parallel cross-industry standard process for text mining (CRISP-TM) is beginning to emerge in the field and is outlined here.

CRISP-TM

CRISP-TM breaks the process of text mining into six major phases (see Figure 2). The arrows in the process diagram indicate the most important and frequent dependencies between phases. The outer circle in the diagram describes the cyclical nature of the text mining process. Lessons learned during one phase can trigger new, often more focused business questions, and subsequent text mining processes will benefit from the experiences of previous ones.

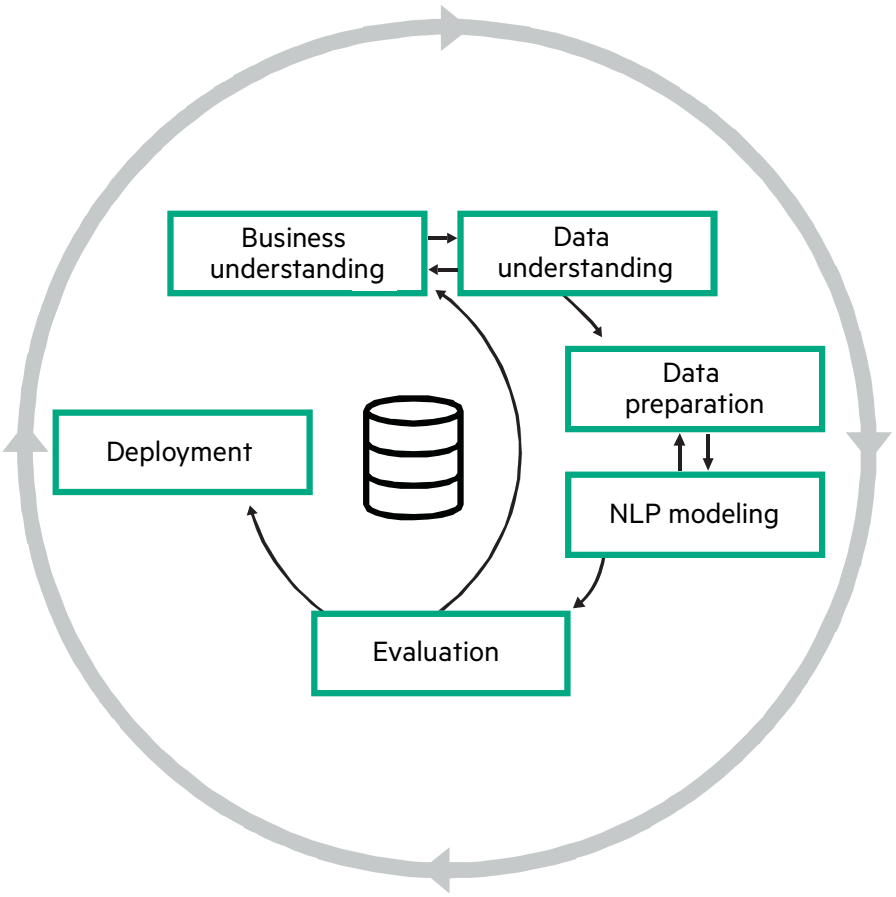


FIGURE 2. Process diagram showing the relationship between the six different phases of CRISP-TM

[2] Text mining, also referred to as text data mining, roughly equivalent to text analytics, is the process of deriving high-quality information from text en.wikipedia.org/wiki/Text_mining.

1. **Business understanding:** This initial phase focuses on understanding the project objectives and requirements from a business perspective. This knowledge can then be converted into a text mining problem definition and a preliminary plan designed to achieve the objectives. A decision model, especially one built using the decision model and notation (DMN) standard, can be used.
2. The **data understanding** phase represents a familiarization with the data sources. During this stage, data quality problems can be uncovered, first insights into the data can be gleaned, and subsets or other structures can be explored to reveal hidden information.
3. At the **data preparation** phase, the final data set is prepared from the initial raw data. Data preparation tasks include table, record, and attribute selection as well as transformation and cleaning of data for modeling tools.
4. In the **NLP modeling** phase, modeling techniques are selected and applied, and their parameters are calibrated. Some techniques have specific requirements in the form of data. Therefore, stepping back to the data preparation phase is often needed.
5. At the **evaluation** stage, the model (or models) are thoroughly assessed to be certain they properly achieve the business objectives. At the end of this phase, a decision on the use of the text mining results should be reached.
6. Creation of the NLP model is generally not the end of the project. Depending on the requirements, the deployment phase can be as simple as generating a report or as complex as implementing a repeatable data scoring (segment allocation) or data mining process. Even if the analyst deploys the model, the customer needs to understand the actions that will be needed to make use of the created models.

3.1 The workflow of an STT solution from edge to core

At the edge, rapid response times are often mandatory. An STT application may generate subtitles quickly, or it may be implemented on an airplane or autonomous vehicle, where delays in system response can represent actual safety hazards. A given piece of speech data may be useful in the core and edge. An optimized STT solution also uses data generated at the edge for longer-term training and AI model refinements performed at the core.

Data generated by sensors at the edge must be quickly analyzed to ensure it's useful enough to send over the network. The core represents the STT system's data center. Today, it may be a physical cluster or based in the cloud, whether private, public, or even distributed between the two.

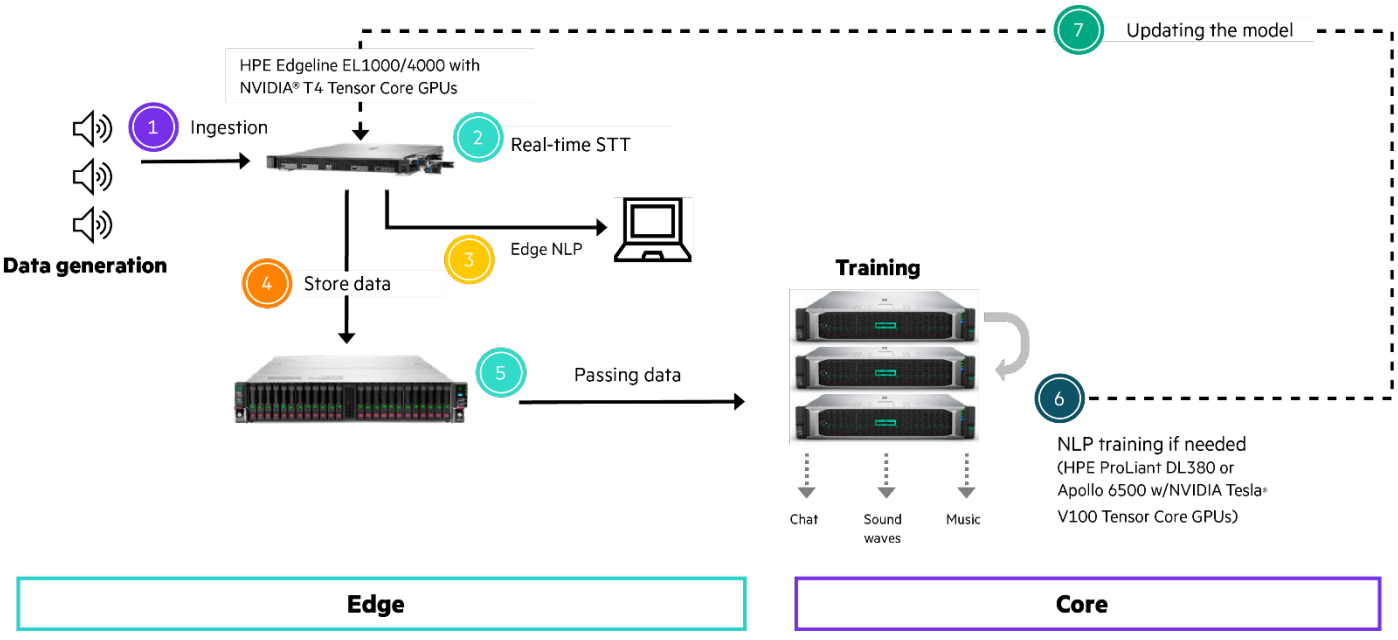


FIGURE 3. STT workflow and components

Figure 2 depicts the workflow of the data from the edge to the core. This workflow more or less applies to all use cases from Table 1. For instance, a sentiment analysis solution in a customer service setting would be implemented as follows:

- **Step 1:** Data is generated during phone conversations with chatbots or human operators.
- **Step 2:** Data is preprocessed and transcribed.
- **Step 3:** The operator can update their decision in real-time, based on the NLP model's input to increase customer satisfaction.
- **Step 4 and 5:** Data is saved either at the call center or at the main data center.
- **Step 6:** Sentiment analysis algorithms are run at the core (data center), discerning customers' emotive responses during the conversation. The newly obtained transcribed data and predictions can be used to further train and improve the model for greater accuracy.
- **Step 7:** Updated sentiment analysis model is deployed at the edge, closer to the operator. Each conversation with a customer will be automatically transcribed and the sentiments inferred.

To create a versatile solution, the data flow needs to be optimized between edge and core. The two complementary directions for data to travel are:

The training/retraining flow:

1. Initial training can be long and requires human intervention to manually associate spoken words with their text equivalents (initial training is typically done well before an STT solution is rolled out) and domain-specific topics require additional training.
2. Retraining happens back in the data center on new data in which the system had previously underperformed. Or perhaps there's new data (typically voice recordings) representing STT examples the system has never encountered before.

The inference flow:

1. Live at the edge: The STT model is running live and producing text as it is spoken and/or generated. Useful immediate action is required and/or live transcription can have immediate value.
2. Analysis at the core: The data being generated at the edge is stored and analyzed later in the data center for a more accurate understanding of the words spoken.

The training/retraining flow is needed to continually improve the model's accuracy. As with any AI model, accuracy increases with the number of pertinent examples that can be used to train the system. Matrix weights are improved and optimized, which can require a substantial amount of computing power, hence a system with multiple GPUs can be very useful.

The inference flow uses the model that has been trained to make analyses and predictions on new data collected at the edge. Of course, the more training and retraining that the system performs, the more accurate the prediction will be.

3.2 Components of an NLP solution

STT is only the beginning of the journey. Other NLP tasks that can then be performed in the workflow after the STT include search, sentiment analysis, and smart transcripts.

An NLP solution, whether STT is involved or not, consists of four primary components:

- Software to perform the STT and/or NLP tasks at hand.
- The training engine typically centered on GPU-based systems at the core.
- Inference engine at the edge.
- Services and support for the solution to maximize customer experience.

HPE has partnered with Intelligent Voice to deliver a tuned and optimized solution for STT that also provides APIs to perform STT and various NLP tasks. These API components are combined to build and solve the different use cases from Table 1:

- High-speed STT, optimized for low-quality and telephony speech (live streaming and batch processing).
- Automatic topic extraction:
  - Quick summary
  - What's trending
  - Semantic navigation

- Smart transcript enhances visualization and navigation of audio files
- Accelerated model building focusses on domain-specific vocabularies
- Biometric search uses voice ID
- Automatic language detection
- Payment card information (PCI) redaction

NLP: Software component

Generic software components: The software for NLP solutions involves an interlocking set of open APIs, enabling the connection between different data flows. Those APIs include STT, NLU, biometric, or other speaker identification, real-time speech and text analysis, ML, and searchable encryption for voice.

Intelligent Voice outputs the text transcript of the audio file with millisecond time stamps and confidence scores for each word. It also produces a lattice of alternative words with the same information.

Domain-specific language model building: Intelligent Voice provides a simple but powerful training process to further enhance the accuracy and usefulness of its STT engine. Its lexicon comes pretuned for common phrases, words, and usage. Jargon and technical terminology particular to specific domains (such as healthcare or financial services) require their own contextualized models.

Intelligent Voice allows textual data input to train the model on unique words, phrases, and acronyms used within a domain. When seeded with contextual domain data (such as existing documents, chat, emails, and Wikipedia pages) Intelligent Voice will extract the new words, add them to the lexicon and retune the algorithms accordingly. In other words, they learn how an organization speaks.

Pairing Intelligent Voice domain-specific vocabulary training with lattice output, the improvement in overall accuracy and retrieval rates can be dramatic. Here is an example of the improved STT accuracy after the model acquires a lexicon specific to pharmaceutical research. Adding the correct input and knowledge from the pharmaceutical world allows the model to identify the correct output.

Before retraining with the domain-specific lexicon, a model could output:

- Kenneth Fisher is chairman and CEO of the pharmaceutical company **merch**. As general **council**, he directed the company's defense against litigation over the anti-inflammatory drug **box**.

After retraining with the correct lexicon and domain knowledge the transcription becomes:

- Kenneth Frazier is chairman and CEO of the pharmaceutical company Merck. As general counsel, he directed the company's defense against litigation over the anti-inflammatory drug Vioxx.

Biometric search and identification: Intelligent Voice's NLP solution augments metadata by capturing a voice profile for every speaker as audio is processed. Each user's voice profile provides another search methodology for users. When reviewing audio, a 30-second snippet can be used to create a search profile for a speaker. A search can then be committed to the database to find every call that potentially contains a particular speaker. Every search result also returns a confidence score indicating the likelihood a given snippet is the searched-for speaker.

Intelligent Voice's biometric identification can be used as part of a multifactor security process, enabling voice profiles as a verification of a person's identity, potentially in conjunction with other security information such as a PIN.

Enrollment can be passive using existing or future-known interactions with a person to create a voice profile from natural conversation. Alternately, enrollment can be formal, asking an individual to speak for a set period to create a voice profile for future use. Within organizations where all voices on a call are profiled, Intelligent Voice's biometric identification software can also identify potential imposters.

Language support and automatic language detection: Intelligent Voice can identify languages spoken, within a larger set of supported languages. It automatically applies the correct language model even if multiple languages are spoken within the same audio file. Any transcript produced will contain the correct language transcription where each language is spoken.

REST-based API: Intelligent Voice has an ecosystem of connectors available out-of-the-box that enables integration with third-party software packages.

NLP and the SmartTranscript: Intelligent Voice has a patent-pending NLP technology that is tuned to deal with difficult, unstructured audio data. Its SmartTranscript feature produces a rich set of semantic data that gives instant insight into what is being said in an audio file. It provides an instant snapshot of any audio or video file to make the content visible without having to listen to a whole file.

The topics that have been extracted from the audio can then be used to create a meta-level view. Using a proprietary topic weighing system, Intelligent Voice promptly surfaces the key topics from a text search of audio. A user can then perform a non-literal search through a data set using hyper-tree technology to link together every audio file with every other one.

Due to Intelligent Voice's flexible ingestion API, any type of text data can be added alongside audio data to give a holistic overview of any data set.

NLP: Hardware component

Core (data center): The data center hosts the core functions of an NLP solution. Its DL capabilities enable the retraining and optimization of your model. These core functions provide the analytics whose compute requirements are too complex to be done at the edge. At the edge, the recording server ingests the audio feeds and sends it to a live archive database. The live database can be placed at the edge or can be sent directly to a central location. As seen in Table 2, HPE has a broad portfolio to host these core functions.

TABLE 2. HPE edge and data center ecosystem for NLP

	Enterprise data center training	Entry data center training and inference		Edge training and inference	Edge inference	
Platform	**Apollo 6500 (XL270d)**	**HPE ProLiant DL380 (XL190r)**	**Apollo 2000 (XL190r)***	**Edgeline EL8000 (e910)**	**Edgeline EL4000 (m710x)**	**Edgeline EL1000 (m710x)**
						
Workload/use case	• Heavy data set training • Mixed training workloads (ex. NLP/Computer vision)	• Light data set training/inference • Small training workloads (ex. NLP)	• Data center inference • High-performance, high density • Up to 2 servers/2U	• Compact high-performance, real-time inference and train at the edge • Rugged environments	• High-performance, real-time inference at the edge • Rugged environments	• Portable high-performance, real-time inference at the edge • Wireless capable • Rugged environments
Acceleration						
Good	Up to 12x NVIDIA T4	Up to 2–4x T4	Up to 2–4x T4	Up to 1x T4**	Up to 1x T4***	Up to 1x T4
Better	Up to 8x Quadro RTX 6000/8000	Up to 5x Quadro RTX 4000	Up to 4x RTX 4000	Up to 8x T4 or 2x RTX 4000/6000**	Up to 2x T4***	Up to 2x T4
Best	Up to 8x NVIDIA Tesla V100-32GB SXM2	Up to 3x V100-32GB PCIe	Up to 2x V100-32GB PCIe	Up to 2x V100-32GB PCIe**	Up to 4x T4***	
	Training focus			Inference focus		
	Data center (scale)				Edge	

* HPE Apollo 2000(XL190r)—acceleration/GPUs is based upon single node configuration.
** HPE ProLiant e910 blade is available in a 1U or 2U tray. Accelerators are installed within each blade.
*** Corresponding ProLiant m710x blade must be configured for each accelerator in the chassis.

Enterprise data center training

HPE Apollo 6500 Gen10 System is an ideal enterprise AI/DL platform that provides performance and flexibility with industry-leading GPUs, fast GPU interconnects, high-bandwidth fabric, and a configurable GPU topology to match varied workloads. The HPE Apollo 6500 System provides reliability, availability, and serviceability (RAS) features. It includes up to eight GPUs per server, next-generation NVIDIA NVLink for fast GPU-to-GPU communication, support for Intel® Xeon® Scalable processors, and a choice of high-speed/low-latency fabric. It is also workload-enhanced using flexible configuration capabilities.

Entry-level data center training and inference

The HPE ProLiant DL380 Gen10 Server is a good starter AI platform for PoCs and small training workloads, delivering the latest in security, performance, and expandability. While backed by a comprehensive warranty, it makes it ideal for any server environment. The HPE ProLiant DL380 Gen10 Server is securely designed to reduce costs and complexity, powered by Intel Xeon Scalable processor, with up to a 60% performance gain[3] and 27% increase[4] in cores, along with HPE DDR4 SmartMemory 2933 MT/s supporting 3.0 TB. It supports 12 Gb/s SAS and up to 20 NVMe drives, plus a broad range of compute options. HPE DC Persistent Memory offers unprecedented levels of performance for databases and analytic workloads. You can run everything from the most basic to mission-critical applications and deploy with confidence.

HPE Apollo 2000 Gen10 System, both for HPC and AI, is great for inference in the data center. It is designed as an enterprise-level, density-optimized, 2U shared infrastructure chassis for up to four HPE ProLiant Gen10 hot-plug servers with all the traditional data center attributes—standard racks and cabling and rear-aisle serviceability access. Server nodes can be serviced without impacting the operation of other nodes in the same chassis to provide increased server uptime. A 42U rack fits up to 20 HPE Apollo 2000 system chassis, accommodating up to 80 servers per rack. It delivers the flexibility to tailor the system to the precise needs of your workload with the right compute, flexible I/O, and storage options. The servers can be mixed and matched within a single chassis to support different applications, and it can even be deployed with a single server, leaving room to scale as customer's needs grow.

Edge training and inference

Ideal for both edge and inference training, the HPE Edgeline EL8000 Converged Edge System is designed to utilize the different compute possibilities to help manage the mounting amount of data being generated. You can deploy and manage HPE Edgeline EL8000, the same way as it were in a traditional data center with HPE ProLiant e910 Server Blades. With a condensed ruggedized form factor, the HPE Edgeline EL8000 is designed for hostile thermal environments of 0°C to 55°C and has a flexible and modular architecture to scale and deliver data center compute capacities and technology directly to where data is created in the world.

Edge inference

The HPE Edgeline portfolio is environmentally ruggedized and supports one-touch provisioning and on-premises/cloud manageability features. The EL1000 is driven by a single Intel Atom®, Core i5/i7, and Xeon E3/Xeon D CPU, respectively; an EL4000 supports four independent Intel Xeon D/E CPUs.

The HPE Edgeline EL4000 Converged Edge System accommodates up to four NVIDIA GPU cards, four independent server cartridges, including the HPE ProLiant m510 Server Blade. This provides it up to 64 Intel Xeon cores, 512 GB memory, up to 16 Terabytes of SSDs, and eight 10GbE ports in a 1U form factor.

The HPE Edgeline EL4000 enables the greatest number of GPUs per rack unit available, offering both high availability and redundancy. It can accommodate up to four NVIDIA Tesla T4 GPUs, each connecting to one server cartridge, providing up to 10240 CUDA cores. Redundant power supplies, ruggedizing (up to MIL-STD through HPE partners), and the backing of industry certifications such as NEBS Level 3, make this a highly reliable system.

Storage

Depending on the amount of input data, the storage solution needed can vary:

- Small storage solution—multiple SDDs in the training server
- Small-to-medium storage solution—network-attached storage varying from HPE MSA based solution to an HPE 3PAR Storage configuration
- Large storage solution—customer call centers running Qumulo or high-performance parallel file systems

[3] HPE measurements: Up to 60% performance increase of Intel Xeon Platinum vs. previous generation E5-2600 v4 average gains of STREAM, LINPACK, SPEC CPU 2006 and SPEC CPU 2017 metrics on HPE servers comparing 2-socket Intel Xeon Platinum 8280 to E5-2699 v4 family processors. Any difference in system hardware or software design or configuration may affect actual performance, conducted in April 2019.

[4] Up to 27% cores increase of Intel Xeon Platinum vs. previous generation comparing 2-socket Intel Xeon Platinum 8280 (28 cores) to E5-2699 v4 (22 cores). Calculation 28 cores/22 cores = 1.27 = 27%, conducted in April 2019.

4. USE CASES

4.1 STT and NLP in law enforcement

Based on HPE AI pilot projects in some law enforcement jurisdictions, more than half of the emergency call center calls are not classified or logged. STT classification of calls plus automated NLP categorization of their content could have a substantial impact on emergency services operations. The functionality of such a solution would include the following:

- Handling prerecorded telephone calls in batches
- Integrating analytics and AI/ML/DL
- Providing anomaly detection and data prediction, impacting workforce demand forecasting
- Using Data Science Studio (Dataiku) to preprocess audio files and related structured and unstructured data sources for transcription and Intelligent Voice used for STT and NLP
- Processing via Intelligent Voice driven by REST API
- Scalable to large volumes

4.2 STT and NLP in the financial services industry

Financial services operate under strict regulatory oversight—leading to sometimes competing and conflicting requirements from compliance departments and IT infrastructure. However, FSI's data size is vast and increasing. No small portion of the world's 175 zettabytes of data generated by 2025[5] could be relevant to FSI market analysis, public sentiment analysis, company analysis, and other FSI analytics. The FSI's data sources include both sliced and legacy systems. The languages represented in these data sets are various. Audio files, especially legacy files (from earlier times when data storage came at a high premium), can be highly compressed. Data sensitivity makes public cloud processing impossible.

The complex combination of factors may make NLP and STT in FSI seem impossible. However, an FSI solution has been developed to provide domain-specific vocabulary training, which improves proactive alerting on complex terms and phrases used by traders in voice, email, and chat. Intelligent Voice's auditable, proactive monitoring solution also satisfies the regulator that appropriate risk management is in place. HPE and Intelligent Voice's comprehensive solution includes STT with multiple language support, smart transcripts, and hyperphonic search.

5. SERVICES

5.1 HPE Artificial Intelligence (AI) Transformation Workshop

Take the first step on your NLP journey with a 1-day HPE Artificial Intelligence Transformation Workshop for key data, business, and IT stakeholders. Depending on your needs and goals, experienced AI and data experts from HPE Pointnext Services will help you:

- Explore use case objectives and priorities for business, data, and IT stakeholders.
- Identify AI and NLP functionalities to reach your objectives.
- Recognize dependencies and data sources to develop an intelligent data strategy.

During the workshop, you will select priority use cases aligned to your business, discover the areas that need attention, and create a high- level plan with opportunities, obstacles, and critical success factors that are specific to your needs. The plan will also include a proof-of-value recommendation to move to an experimentation stage with your data and in your environment.

5.2 HPE AI Agile Design and Planning Service

With HPE AI Agile Design and Planning Service, following a series of workshops and interview sessions, senior AI and data experts from HPE Pointnext Services will facilitate your AI initiative implementation. The choice of project management methodologies, based on the HPE Pointnext Services experience and knowledge, is driven by the need to:

- Coordinate, communicate, and collaborate while adopting innovation and overtaking analytic silos.
- Apply an iterative and incremental approach to remain within budget allocation boundaries.
- Tailor the solution for the scope of your solution's architecture and design.

[5] The Digitization of the World From Edge to Core, IDC, November 2018

5.3 Consumption-based IT services

Gain the flexibility and control of the on-premises public cloud with HPE GreenLake—a set of consumption-based IT solutions. Choose from a catalog of complete, curated solutions that deliver IT outcomes with hardware, software, and expertise on-premises in a pay-per-use model.

HPE Pointnext Services will implement and operate these solutions for you, enabling you to focus your own IT resources where they add the most business value. Or you can consume the technology of your choice, also using the pay-per-use model, in a manner suited to how you operate IT. The consumption models include:

- Preconfigured, end-to-end HPE GreenLake Solutions that deliver the time-to-value for outcomes such as Apache Hadoop, backup, edge compute, open-source database platform, and SAP HANA®.
- Modular HPE GreenLake solutions offer infrastructure choice such as containers, Microsoft Azure Stack, storage, VMs, high-performance computing (HPC), or other technologies from HPE.

5.4 Operational Support Services

HPE redefined the concept of operational efficiency helping to create new IT experiences for our customers' business, from the core to the edge. HPE offers a wide range of predefined, packaged support services, which customers can select and tailor, from the following options:

- HPE Foundation Care is a cost-effective support to help customers when there is a problem.
- HPE Proactive Care builds on HPE Foundation Care but provides an Enhanced Call Experience in case of a failure.
- HPE Proactive Care Advanced includes all the deliverables of HPE Proactive Care and includes a dedicated resource for more personalized support.
- HPE Datacenter Care is a combination of all support services for all infrastructure components into a single support contract.

6. EXAMPLE CONFIGURATION

HPE together with Intelligent Voice has been testing and packaging the best configurations suited for a variety of customer needs. Some predefined configurations are:

Data center

- Light and medium training data (the amount of data is smaller and the number of accelerators needed is not high): These system needs are best addressed by the HPE ProLiant DL380 Gen10 with two Intel Xeon Gold processors, at least 128 GB of memory and 2 x 400 GB disks. (The storage solution should be increased depending on the amount of data and data strategy.) The system should also be equipped with 2 NVIDIA V100 GPUs.
- Heavy training data: In this case, the HPE Apollo 6500 Gen10 is equipped with 4 or 8 NVIDIA V100 GPUs.

Edge

- HPE Edgeline EL1000 can host an HPE ProLiant m710x server, which runs the live STT depending on the number of channels that are being fed in parallel. An NVIDIA T4 GPU can also be added to the configuration if needed.
- HPE Edgeline EL4000 can host up to 4 x HPE ProLiant m710x cartridges, which can normally run STT transcription in parallel. Depending on the number of channels the system hosts, GPU accelerators may also be added.
- HPE Edgeline EL8000 is a very powerful edge device, which has been built for both running live inference and retraining at the edge. It can also be equipped with the latest NVIDIA V100 GPUs.

Software

- HPE supports various operating systems such as RHEL and SLES
- The installation is normally done on bare metal but can be containers driven
- Depending on the number of servers the deployment can be done manually or driven by our cluster manager HPE PCM
- Intelligent Voice software

Storage

- It can range from the internal-attached storage array to network-attached storage to higher-performance solutions such as parallel file systems.

7. SUMMARY

To perform world-class NLP and STT operations, world-class IT is essential. Hewlett Packard Enterprise offers you a strategic advantage with expertise, technology, and partnerships. Years of delivery expertise across technology, geographies, and AI workloads distinguish HPE's subject-matter experts, including data scientists, AI presales teams, solution architects, centers of excellence, and AI benchmarking engineers. Integrated AI solutions from HPE encompass servers, storage, software, networking, and services that power the full end-to-end workflow. Our partners are proven ISVs, system integrators, channel partners and distributors, technology partners, and service providers.

The NLP market is still developing and is expected to expand significantly until at least 2025. NLP's rapid growth will be powered in part by affordable and scalable infrastructure. A major portion of the expanding NLP market is driven by services that enable organizations to adopt and leverage the power of NLP. Industries such as retail, banking, and utilities are expected to be early NLP adopters. Human-computer interaction is another new, multi-industry frontier in which NLP and STT solutions will surely play an important role.

Contact your HPE account manager for more information, reference configurations, use cases, service, and support offerings, along with the latest HPE and partner technology for world-class NLP solutions.

LEARN MORE AT

Navigate AI and realize what it can offer, know more about having HPE as your trusted AI partner

hpe.com/ai

hpe.com/info/deep-learning

Follow us on
Facebook: facebook.com/HPEAI/
Twitter: HPE AI - @HPE_AI
LinkedIn: linkedin.com/showcase/hpe-ai/

a00091770ENW, December 2019

Hewlett Packard Enterprise