Making Sense of Sounds, or How Analog Devices' AI Can Boost Your Machine Uptime

By **Sebastien Christian**

Introduction

Anyone familiar with the necessity of maintaining a mechanical machine knows how important the sounds and vibrations it makes are. Proper machine health monitoring through sound and vibrations can cut maintenance costs in half and double the lifetime. Implementing live acoustic data and analysis is another important approach for condition-based monitoring (CbM) systems.

We can learn what the normal sound of a machine is. When the sound changes, we identify it as abnormal. Then we may learn what the problem is so that we can associate that sound with a specific issue. Identifying anomalies takes a few minutes of training, but connecting sounds, vibrations, and their causes to perform diagnostics can take a lifetime. There are experienced technicians and engineers with this knowledge, but they are a scarce resource. Instinctively recognizing a problem from sound alone can be difficult, even with recordings, descriptive frameworks, or in-person training with experts.

Because of this, our team at Analog Devices has spent the last 20 years on understanding how humans make sense of sounds and vibrations. Our objective was to build a system able to learn sounds and vibrations from a machine and decipher their meaning to detect abnormal behavior and to perform diagnostics. This article details the architecture of OtoSense, a machine health monitoring system that enables what we call computer hearing, which allows a computer to make sense of the leading indicators of a machine's behavior: sound and vibration.

This system applies to any machine and works in real time with no network connection needed. It has been adapted for industrial applications and it enables a scalable, efficient machine health monitoring system.

This article delves into the principles that guided OtoSense's development, and the role of human hearing in designing OtoSense. The article then discusses the way sound or vibration features were designed, how meaning is derived from them, and the continuous learning process that makes OtoSense evolve and improve over time to perform increasingly complex diagnostics with increasing accuracy.

Guiding Principles

To be robust, agnostic, and efficient, the OtoSense design philosophy followed some guiding principles:

- **Get inspiration from human's neurology.** Humans can learn and make sense of any sound they can hear in a very energy efficient manner.
- **Be able to learn stationary sounds as well as transient sounds**. This requires adapted features and continuous monitoring.
- **Perform the recognition at the edge, close to the sensor**. There should not be any need of a network connection to a remote server to make a decision.
- **Interaction with experts and the necessity to learn from them must happen with minimal impact on their daily workload**, and be as enjoyable as possible.

The Human Hearing System and Translation to OtoSense

Hearing is the sense of survival. It's the holistic sense of distant, unseen events, and it matures before birth.

The process by which we humans make sense of sounds can be described in four familiar steps: analog acquisition of the sound, digital conversion, feature extraction, and interpretation. In each step, we will compare the human ear with the OtoSense system.

- **Analog acquisition and digitization**. A membrane and levers in the middle ear capture sounds and adjust impedance to transmit vibrations to a liquid-filled canal where another membrane is selectively displaced depending on spectral components present in the signal. This in turn bends flexible cells that emit a digital output that reflects the amount and harshness of the bending. These individual signals then travel on parallel nerves arranged by frequency to the primary auditory cortex.
  - In OtoSense, this job is performed by sensors, amplifiers, and codecs. The digitization process uses a fixed sample rate adjustable between 250 Hz and 196 kHz, with the waveform being coded on 16 bits and stored on buffers that range from 128 samples to 4096 samples.
- **Feature extraction** happens in this primary cortex: Frequency-domain features such as dominant frequencies, harmonicity, and spectral shape, as well as time-domain features such as impulsions, variations of intensity, and main frequency components over a time window spanned around 3 seconds.
  - OtoSense uses a time window that we call chunk, which moves with a fixed step size. The size and step of this chunk can range from 23 ms to 3 s, depending on the events that need to be recognized and the sample rate, with features being extracted at the edge. We'll provide more information on the features extracted by OtoSense in the next section.

- **Interpretation happens in the associative cortex**, which merges all perceptions and memories and attaches meaning to sounds, such as with language, which plays a central role in shaping our perceptions. The interpretation process organizes our description of events far beyond the simple capacity of naming them. Having a name for an item, a sound, or an occurrence allows us to grant it greater, more multilayered meaning. For experts, names and meaning allow them to better make sense of their environment.
  - This is why OtoSense interaction with people starts from visual, unsupervised sound mapping based on human neurology. OtoSense shows a graphical representation of all the sounds or vibration heard, organized by similarity, but without trying to create rigid categories. This lets experts organize and name the groupings seen on screen without trying to artificially create bounded categories. They can build a semantic map aligned with their knowledge, perceptions, and expectations regarding the final output of OtoSense. The same soundscape would be divided, organized, and labelled differently by auto mechanics, aerospace engineers, or cold forging press specialists—or even by people in the same field but at different companies. OtoSense uses the same bottom-up approach to meaning creation that shapes our use of language.

From Sound and Vibration to Features

A feature is assigned an individual number to describe a given attribute/quality of a sound or vibration over a period of time (the time window, or chunk, as we mentioned earlier). The OtoSense platform's principles for choosing a feature are as follows:

- **Features should describe the environment** as completely as possible and with as many details as possible, both in the frequency domain and time domain. They have to describe stationary hums as well as clicks, rattles, squeaks, and any kind of transient instability.
- **Features should constitute a set as orthogonally as possible**. If one feature is defined as "the average amplitude on the chunk," there should not be another feature strongly correlated with it, as a feature such as "total spectral energy on the chunk" would be. Of course, orthogonality is never reached, but no feature should be expressed as a combination of the others—some singular information must be contained in each feature.
- **Features should minimize computation**. Our brain just knows addition, comparison, and resetting to 0. Most OtoSense features have been designed to be incremental so that each new sample modifies the feature with a simple operation, with no need for recomputing it on a full buffer or, worse, chunk. Minimizing computation also implies not caring about standard physical units. For example, there is no point in trying to represent intensities with a value in dBA. If there is a need to output a dBA value, it can be done at the time of output if necessary.

A portion of the OtoSense platform's two to 1024 features describe the time domain. They are extracted either right from the waveform or from the evolution of any other feature over the chunk. Some of these features include the average and maximal amplitude, complexity derived from the linear length of the waveform, amplitude variation, the existence and characterization of impulsions, stability as the resemblance between the first and last buffer, skinny autocorrelation avoiding convolution, or variations of the main spectral peaks.

The features used on the frequency domain are extracted from an FFT. The FFT is computed on each buffer and yields 128 to 2048 individual frequency contributions. The process then creates a vector with the desired number of dimensions—much smaller than the FFT size, of course, but that still extensively describe the environment. OtoSense initially starts with an agnostic method for creating equal-sized buckets on the log spectrum. Then, depending on the environment and the events to be identified, these buckets adapt to focus on areas of the spectrum where information density is high, either from an unsupervised perspective that maximizes entropy or from a semi-supervised perspective that uses labelled events as a guide. This mimics the architecture of our inner ear cells, which is denser where the speech information is maximal.

Architecture: Power to the Edge and Data on Premises

Outlier detection and event recognition with OtoSense happen at the edge, without the participation of any remote asset. This architecture ensures that the system won't be impacted by a network failure and it avoids having to send all raw data chunks out for analysis. An edge device running OtoSense is a self-contained system describing the behavior of the machine it's listening to in real time.

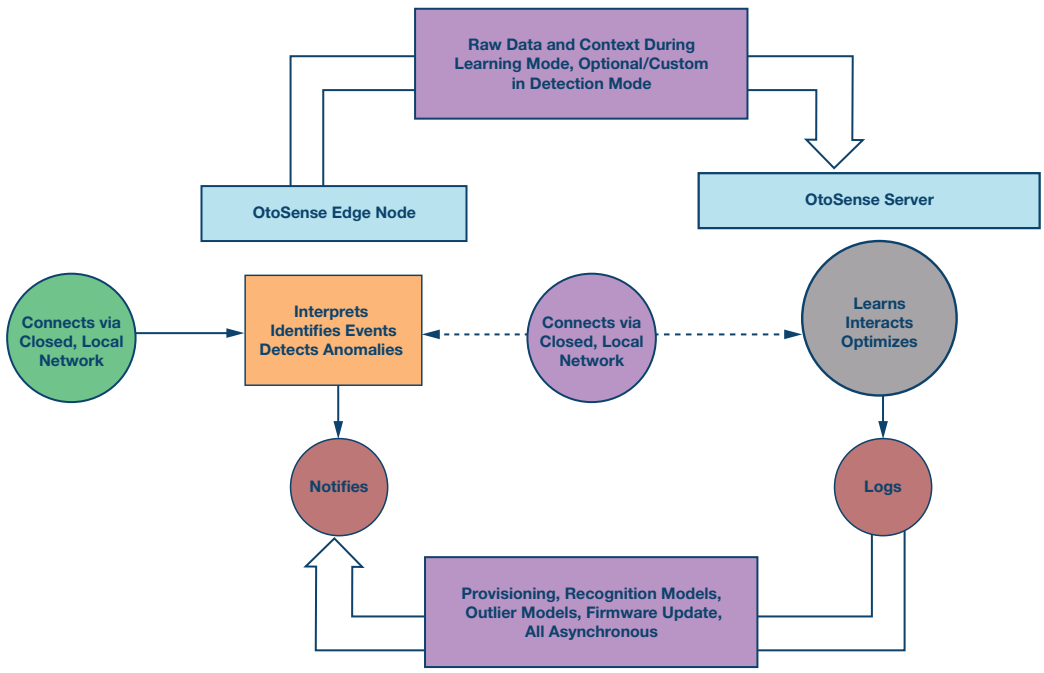

Figure 1. The OtoSense system.

The OtoSense server, running the AI and HMI, is typically hosted on premises. A cloud architecture makes sense for aggregating multiple meaningful data streams as the output of OtoSense devices. It makes less sense to use cloud hosting for an AI dedicated to processing large amounts of data and interacting with hundreds of devices on a single site.

From Features to Anomaly Detection

Normality/abnormality evaluation does not require much interaction with experts to be started. Experts only need to help establish a baseline for a machine's normal sounds and vibrations. This baseline is then translated into an outlier model on the Otosense server before being pushed to the device.

We then use two different strategies to evaluate the normality of an incoming sound or vibration:

- The first strategy is what we call usualness, where any new incoming sound that lands in the feature space is checked for its surrounding, how far it is from baseline points and clusters, and how big those clusters are. The bigger the distance and the smaller the clusters, the more unusual the new sound is and the higher its outlier score is. When this outlier score is above a threshold as defined by experts, the corresponding chunk is labelled unusual and sent to the server to become available for experts.
- The second strategy is very simple: any incoming chunk with a feature value above or below the maximum or minimum of all the features defining the baseline is labelled as extreme and sent to the server as well.

The combination of unusual and extreme strategies offers good coverage of abnormal sounds or vibrations, and these strategies perform well for detecting progressive wear and unexpected, brutal events.

From Features to Event Recognition

Features belong to the physical realm, while meaning belongs to human cognition. To associate features with meaning, interaction between OtoSense AI and human experts is needed. A lot of time has been spent following our customers' feedback to develop a human-machine interface (HMI) that enable engineers to efficiently interact with OtoSense to design event recognition models. This HMI allows for exploring data, labelling it, creating outlier models and sound recognition models, and testing those models.

The OtoSense Sound Platter (also known as splatter) allows for the exploration and tagging of sounds with a complete overview of the data set. Splatter makes a selection of the most interesting and representative sounds in a complete data set and displays them as a 2D similarity map that mixes labelled and unlabelled sounds.

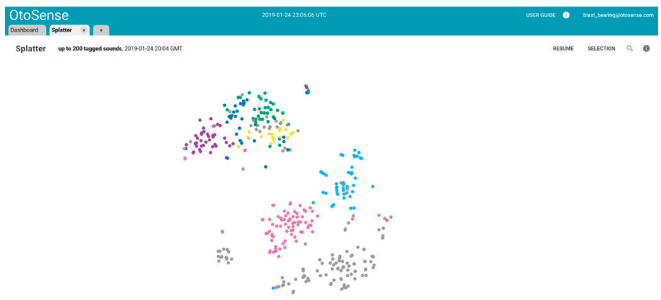

Figure 2. A 2D splatter map of sound in the OtoSense Sound Platter.

Any sound or vibration can be visualized, along with its context, in many different ways—for example, using Sound Widgets (also known as Swidgets).

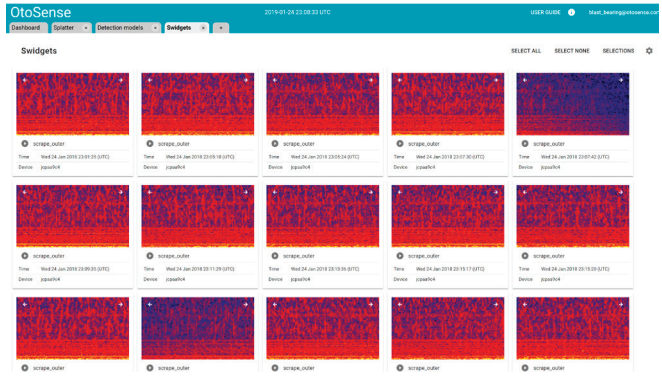

Figure 3. OtoSense sound widgets (swidgets).

At any moment, an outlier model or an event recognition model can be created. Event recognition models are presented as a round confusion matrix that allows OtoSense users to explore confusion events.

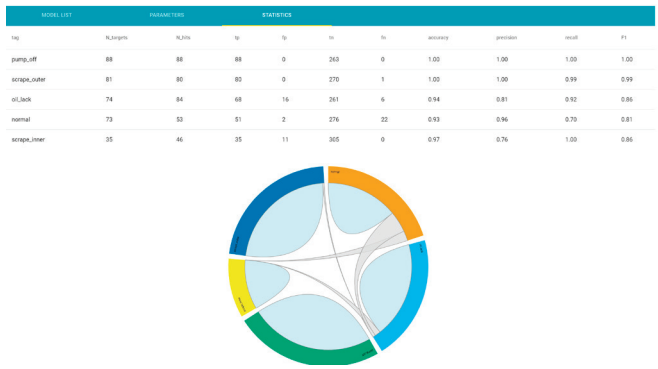

Figure 4. An event recognition model can be created based on the required events.

Outliers can be explored and labelled through an interface that shows all the unusual and extreme sounds over time.

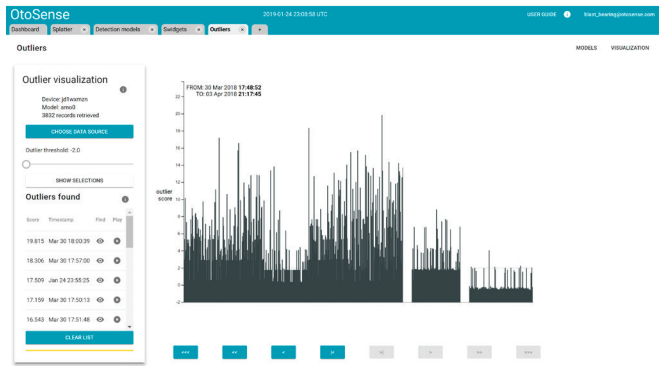

Figure 5. Sound analytics over time in the OtoSense Outlier visualization.

The Continuous Learning Process, from Anomaly Detection to Increasingly Complex Diagnostics

OtoSense has been designed to learn from multiple experts and allow for more and more complex diagnostics over time. The usual process is a recurring loop between OtoSense and experts:

- An outlier model and an event recognition model are running at the edge. These create output for the probability of potential events happening, along with their outlier scores.
- An unusual sound or vibration above the defined threshold triggers an outlier notification. Technicians and engineers using OtoSense can then check on the sound and its context.

- These experts then label this unusual event.
- A new recognition model and outlier model that includes this new information is computed and pushed to edge devices.

Conclusion

The objective of the OtoSense technology from Analog Devices is to make sound and vibration expertise available continuously, on any machine, with no need for a network connection to perform outlier detection and event recognition. This techonology's growing use for machine health monitoring in aerospace, automotive, and industrial monitoring applications has shown good performance in situations that once require human expertise and in situations involving embedded applications, especially on complex machines.

References

Christian, Sebastien. "How Words Create Worlds." TEDxCambridge, 2014.

Sebastien Chistian [sebastien.christian@analog.com] had an early passion for understanding how we humans build an inner, sharable model of the world, using our senses, and how we use this model to describe the world they live in.

Sebastian earned an M.S. in quantum physics, which he followed with an M.S. in neuroscience and a third degree in semantics. Sebastien's education combined research, development, and field experiments. He worked as a speech and language pathologist with psychotic and deaf children for 10 years, refining his understanding of sensor-based meaning creation and sharing, with an emphasis on hearing. Sebastien says that this practice, where he worked with the same young patients for years, is what brought all the scattered pieces of knowledge together into a single, coherent picture.

During the same period, Sebastien became an expert for the French Ministry of Health, where he advised on hearing loss policies, taught in medical school and at Paris Sorbonne University, and, in 2011, created the first independent private R&D laboratory dedicated to bringing AI driven innovations to people with sensing and cognitive disabilities.

In 2013, Sebastien completed a full prototype of his machine hearing project, which earned him laureate of the NETVA tech competition in Cambridge, MA. The massively positive feedback from his fellows at MIT and from the business community led him to found OtoSense in early 2014, and to develop what is the first AI focused on making sense of any sound. This machine hearing platform revealed itself to be well adapted to complex environments and complex machine monitoring.

After receiving multiple awards, which included a Best App of the Year award at GSMA Mobile World Congress in 2015, OtoSense focused on machine monitoring in the industrial and transportation verticals, with vast and growing range of potential applications ahead.

Sebastien is now leading OtoSense inside product development at Analog Devices.

Sebastien Christian