ConVersa by Red Box

The Secrets to Maximising your Voice & AI Strategy

Why it's Time to Rethink Your Recording

IT'S TIME TO RETHINK YOUR RECORDING

CONTENTS

The advanced capabilities of speech analytics and associated use cases are making organisations wake up to the value held in their call recordings. However, despite the potentially game changing ability to tap into the rich insights from the analysis of audio data at scale, challenges remain.

Established call recording practices that have served organisations well enough over the years for compliance and quality purposes are now hindering the ROI from speech analytics.

This Whitepaper explains why you will need to rethink some of the standard recording practices that were established long before AI-powered voice analytics was a possibility, empowering you to gain a competitive advantage and see the real benefits from a voice and AI strategy.

THE TRUTH ABOUT CALL RECORDING

The widespread adoption of digital devices, an increasing desire for personalisation and demand for quality of service is changing the way consumers want to communicate with organisations.

In today's Contact Centre, interactions are being transformed by omni-channel engagements that allow customers to interact with agents seamlessly. A transition driven largely by organisations who want to gain operational efficiencies, reduce costs, and identify cross-sell and upsell opportunities to improve the customer experience, call recording in particular is a key element to this success.

FUELLING WORKPLACE TRANSFORMATION

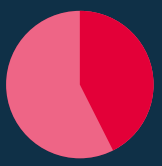
Not only is call recording now critical for many organisations where regulatory requirements are concerned, but it is being leveraged to foster a much broader sense of digital enablement. This includes using valuable call data to fuel AI and drive differentiation as part of digital transformation, whether it be feeding the data via open APIs for business intelligence and customer insights, or using voice transcription to monitor communications and identify opportunities.

> "AI is being leveraged to foster a much broader sense of digital enablement"

THE CURRENT STATE OF PLAY

Today, the vast majority of Contact Centres have recording in place already. However, the reality is that traditional recording practices that provide call audio for the human ear to listen to and assess are simply not good enough when feeding AI engines with huge volumes of audio data. In fact, accessing the data is one thing, but getting quality data in a structured format and timely manner to gain valuable real-time insights is another.

So, where do you start in order to maximise your voice and AI strategy? It's time to rethink your recording.



38%

OF CALL CENTRES HAVE VOICE ANALYTICS ON THEIR WISH LIST*

*According to the 2020 Call Centre Helper Survey

AI IS CHANGING HOW VOICE IS BEING USED

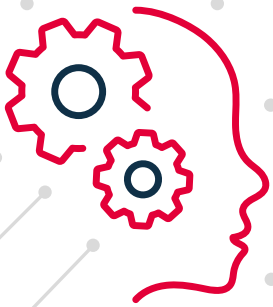
Voice is a uniquely rich data set that, when combined with AI, delivers powerful insights and intelligence to drive true and measurable business outcomes.

Voice analytics has come of age, with a growing technology landscape of AI vendors finding new and improved ways to analyse complex audio datasets. These tools can surface real-time, actionable insights that were not possible a few years ago. Timely access to these insights and trust in the underlying data driving them can enable fast and agile innovation in increasingly competitive and disrupted markets.

"Timely access to AI-driven insights and trust in the underlying data driving them can enable fast and agile innovation in increasingly competitive and disrupted markets."

INVALUABLE AI INSIGHTS PRESENTING ENDLESS OPPORTUNITIES FOR ORGANISATIONS:

- Fraud Detection
- Trader Oversight
- Risk Management
- Behavioural Detection
- Automation
- Supply Chain



- Customer Experience
- Sales Performance Optimisation
- Predictive Analytics
- Employee Engagement
- Hyper-personalisation
- Customer Support

QUALITY IN = QUALITY OUT

As AI analytics emerge into the mainstream, it is becoming clear in many instances that the Automatic Speech Recognition engines are not receiving the quality of voice data to be effective. So how much of a problem is this?

To illustrate the impact of audio quality on the accuracy of transcription, take a look at the following table which shows two transcriptions using a G.711 Lossless Codec and a G.729a Lossy Codec. An explanation of the Lossless and Lossy terms will be given later in the Whitepaper.

TEST SCRIPT:	G.711 TRANSCRIPT:	G.729A TRANSCRIPT:
This is a test call from Red Box Recorders. One two three one two three.	This is a test call from Red Box Recorders. One two three one two three please.	I'll just call her at Fox record. One two three one two three please.
Please buy 200 shares at 14 dollars per share. Simon sells seashells by the seashore.	By two hundred shares at fourteen dollars per share. Simon sales seashells by the seashore.	<miss> Five hundred <miss> of forking dollars per share. What else. Five as she shared
Test call will end now. Three two one Goodbye.	Tesco will end up three to one to five.	Tough call. Three two one two guys.
	Words: 41, Errors: 2.5. Accuracy: 94%	Words: 41, Errors: 16. Accuracy: 61%

As you can see, the end result is a transcription that is accurate enough to provide a useful insight into what was said and a transcription that loses much of its sense. Now, imagine that repeated for every call you record. What useful analysis could even the best AI system perform on that?

What this illustrates is that good transcriptions must be based on giving the transcription software the highest quality of audio possible.

COMPUTERS ARE NOT AS GOOD AT UNDERSTANDING SPEECH AS HUMAN BEINGS

Until voice transcription became possible, recording was a tool primarily accessed by the human ear. Even the main quality metric of many telephony systems has been driven by the subjective assessment of the human ear.

The “Mean Opinion Score” (MoS), is calculated as the arithmetic mean over single ratings performed by human beings. Judged on a scale of 1 (bad) to 5 (excellent), it is a common metric for a voice quality measurement in telephony.

For the two most commonly used Codecs **G711** and **G729a**, the MoS are around 4.1-4 and 3.7-4 respectively, depending on the environment. These might sound close but look at the example above. It is the difference between a useful transcription and a useless one because it is a measure that assumes human listening.

In truth, humans are very good at making sense of speech under very difficult circumstances. We hold conversations in noisy restaurants and on trains with lots of background noise. We can ignore extraneous or irrelevant sounds and hold intelligible conversations.

Computers are not yet this sophisticated and transcription services need the best audio quality they can get to be effective. So, if you want to get the most out of your analytics application, you need to address this reality.

WHAT ARE CODECS & WHY DO THEY MAKE A DIFFERENCE?

Codecs determine how voice is carried across a network and how it is stored.

A device or program that prepares calls for transmission or storage & enables them to be understood at its destination.

A mashup of two words – “encode” & “decode”, an audio data stream is encoded for transmission or storage & decoded for playback.

Codecs decide size of voice data for both storage and the amount of traffic or bandwidth needed to transfer the data across a network.

THEY DO THIS IN THREE WAYS:

1. SAMPLING RATE

- **Sampled bit rate per second:** This is effectively the amount of data that is captured per second from a call, which in turn reflects the granularity of a call recording in the way that a megapixel captures the granularity in a picture. In the example above G.711 captures data at a rate of 64kbits/s and G.729a is captured at a considerably less granular 8kbits/s.

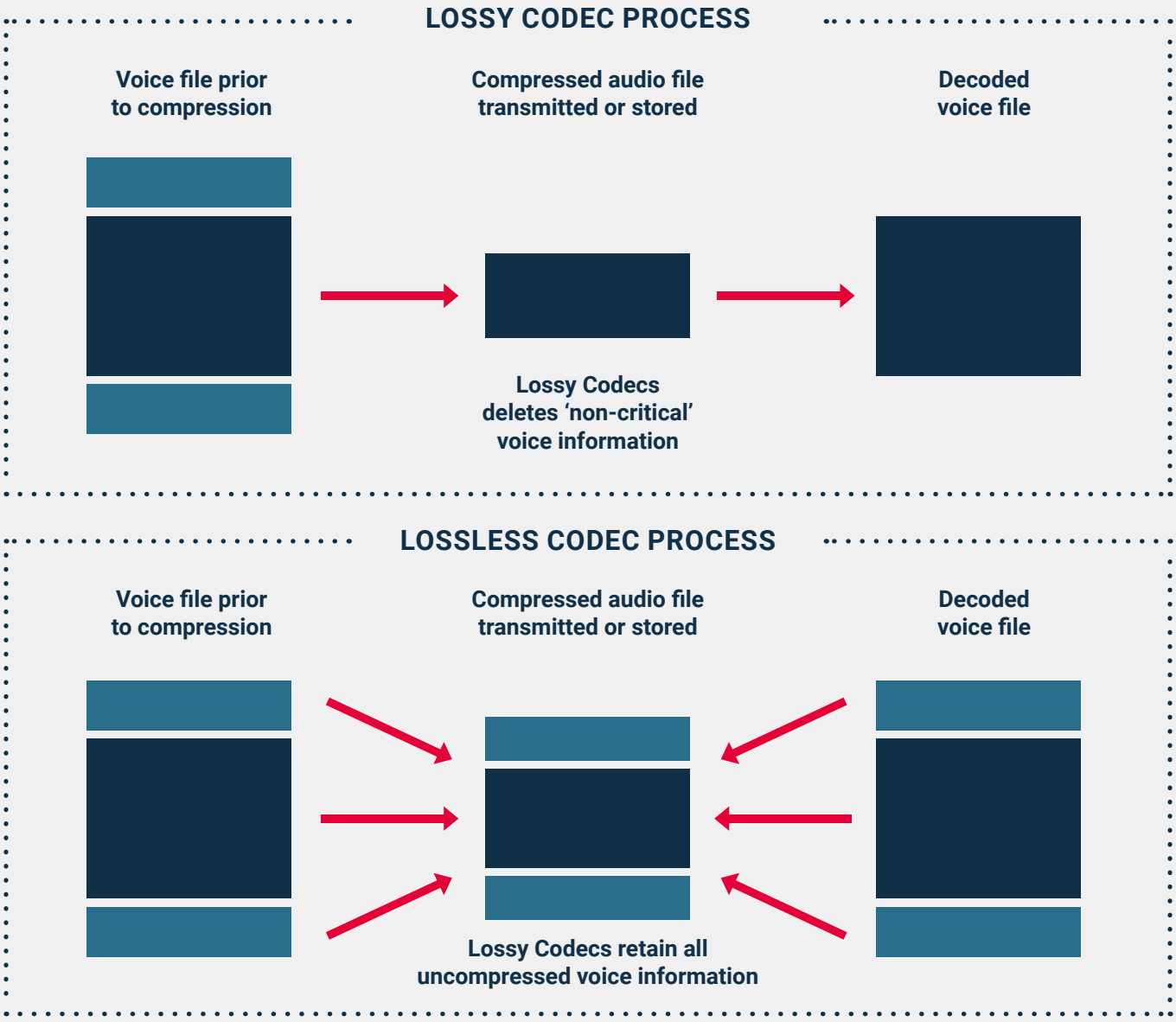
  So, a G.711 Codec captures much greater detail for a transcription engine to focus on. However, the file size is resultantly bigger. G.711 effectively captures uncompressed data at the quality of the original voice call, whereas G.729a captures a reduced quality call.

CODEC	COMPRESSES CALL RECORDINGS	REDUCES RECORDING QUALITY	BITRATE PER SECOND	STORAGE REQUIREMENTS	BANDWIDTH REQUIREMENTS
Lossy - G.729	Yes	Yes	8 Kbps	Relatively low	Relatively low
Lossless - G.711	No	No	64 Kbps	Higher	Higher

2. COMPRESSION

- **Codecs** will then compress the call data on transmission or when it is stored. It will then decompress the files for transmission or playback. However, with some Codecs the files will lose some of the information held in the original file. This has given rise to Codecs being divided in to “Lossy” and “Lossless” Codecs.

- **Lossless Codecs** exploit redundant data and so when the file is decompressed, the audio file has lost none of its granularity. But with a Lossy Codec the process is more complicated, effectively exploiting not just redundant data, but also human perception. The Lossy Codec processes the voice file and sends a much more highly synthesised version of the voice data instead of a point-by-point recreation of the waveform. A Lossy Codec will try to keep the core elements of a voice file but will reject elements not seen as key. Put simply, the algorithms that do this were developed in an age when meeting the needs of the human ear was good enough!



G.729 is the Codec favoured by VoIP providers as a "bandwidth saver", so that they can connect slower connections. G.729 compresses 64Kbps into only 8Kbs, a compression ratio of 8 to 1, but in practice you get bandwidth savings of about 3 or 4 to 1 when RTP and packet headers are considered. Whereas G.711 compression is closer to 50%. G.729 is able to transmit voice very efficiently–at about 32 kBit/s versus 87 kBit/s for G.711.

- **Lossy Codecs** enable greater compression and so have real benefits when it comes to bandwidth in a network or storage. This can enable greater call volumes to be handled more easily during busy times. However, the impact on transcription can render the result effectively useless as seen in the above example.

- Some studies have also shown that Lossy Codecs can have distinct effects on timbral and emotional characteristics, in some cases strengthening negative emotional qualities and weakening positive ones. As voice analytics looks to capture the emotions of calls, this too will have negative effect on the effectiveness of analytics.

- To compound the issue, Lossy Codecs lose detail every time a Codec is applied to an audio file and when Adaptive Network Codecs are used to manage network traffic, detail may be lost several times in a complex network.

- Lossy Codecs are something that you need to be aware of if a recording vendor is using their own proprietary Codec, which some vendors do.

> Lossy Codecs can have distinct effects on timbral and emotional characteristics, in some cases strengthening negative emotional qualities

3. FREQUENCY

- **Frequency:** VoIP telephony and recording systems do not record every sound that a human ear can hear. They will focus on the frequencies that are typically used by the human voice. Humans can typically hear frequencies in the range of 20 Hz and 20,000 Hz. But the usable frequency range of a voice telephony system is typically in the range of 300 to 3400 Hz and Codecs like G.711 and G.729a will match this range.

  Some Codecs like G.722 will extend the frequency range from as low as 50Hz to 7000Hz. By capturing a wider frequency range, the recordings can hold a greater range of sound information. This doesn't impact transcription to a great extent because most of the extra frequency range captured includes sounds outside the range commonly used by the human voice.

> Humans can typically hear frequencies in the range of **20 Hz** and **20,000 Hz.** But the usable frequency range of a voice telephony system is typically in the range of **300** to **3400 Hz**

STEREO VS. MONO

IS YOUR TRANSCRIPTION EFFECTIVE?

CAN YOU IDENTIFY WHAT IS BEING SAID AND WHO IS SAYING IT?

With stereo recording this is simple because each channel is recorded separately. It also has an advantage because background noise from one channel that might affect transcription WER accuracy will not affect the other. And of course, where there is overlapping speech which is a common occurrence.

IS DIARIZATION THE SOLUTION?

One potential solution that is offered is Diarization – a method whereby the application will identify different speakers. Diarization works by dividing the mono recording into small samples, detecting which contain speech and which do not (Speech Activity Detection – SAD), and then using algorithms to identify similar patterns so that speakers can be identified. Whilst results may be impressive at a technical level it is by no means the finished article, with research striving to improve the typical Diarization Error Rate (DER).

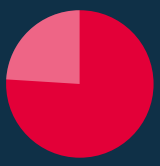
In addition to the technical challenges of the process itself, things like background noise on one side of the call when both speakers are talking, or simply overlapping speech in a heated exchange, add to these challenges. Again, this is another element that impacts the speech-to-text accuracy of transcription.

In summary, your starting point should be stereo recording.

REAL-TIME AND OPEN ACCESS TO HIGH-QUALITY METADATA

Demand to turn valuable data into actionable insights is growing

With more value than any other means of communication given its ability to convey sentiment and context, demand for accurate, real-time access to voice data to fuel customer analysis, increase efficient operations and enhance competitive advantage is higher than ever.



76%

of Senior IT Executives regard voice as valuable or very valuable to their organisation*

*Survey conducted by SAPIO Research for Red Box

METADATA IS KEY

A crucial element to this success is call metadata - containers of information that describe the contents and context of data collected from call recordings, such as time, duration, the caller's phone number, agent ID and the call journey.

To ensure you receive the optimal metadata possible, this should be captured via a CTI connector. Metadata is collected when calls are recorded digitally and integrated into PBXs and other telephony systems, allowing the server to capture information that can be 'tagged' to the call.

Once you have the metadata, you have the control, and knowledge to engineer, report and present data however you choose. You can gather information for every recorded call for any agent in your organisation to establish:

- ARE THEY MEETING KPI TARGETS?
- IS YOUR ORGANISATION MEETING SLAS?
- ARE CUSTOMER QUERIES BEING ANSWERED IN A TIMELY MANNER, OR DOES THE CALL FLOW NEED AMENDING?

METADATA USE CASE:

A client raises a complaint that their call hasn't been answered quickly enough and they wanted to speak to finance, however they keyed the wrong option from the automated system and get connected to technical support.

The technical support team place the customer back into the call queue for the finance team and the client is held in the queue for a few minutes before speaking to the right person.

Metadata captures every step of the call flow – the number the client dialled, the number keyed on the automated system to be answered by the technical support team, and the client being placed back into a call queue before speaking to the finance department.

METADATA TO FUEL AI

It's also important that organisations continuously review their metadata as their goals evolve over time, as refining this information translates into more valuable call recordings that can also be enhanced through the likes of speech analytics and AI applications. For example, speech analytics can create infinite tags, identifying and instantaneously retrieving calls by specific words or phrases such as the name of a competitor or statements like: "I am not satisfied with the service you're providing". Moreover, speech analytics will take authorised personnel directly to the point of the call where that word or phrase is mentioned, preventing them from having to search through the whole call for the information they need for a more efficient process.

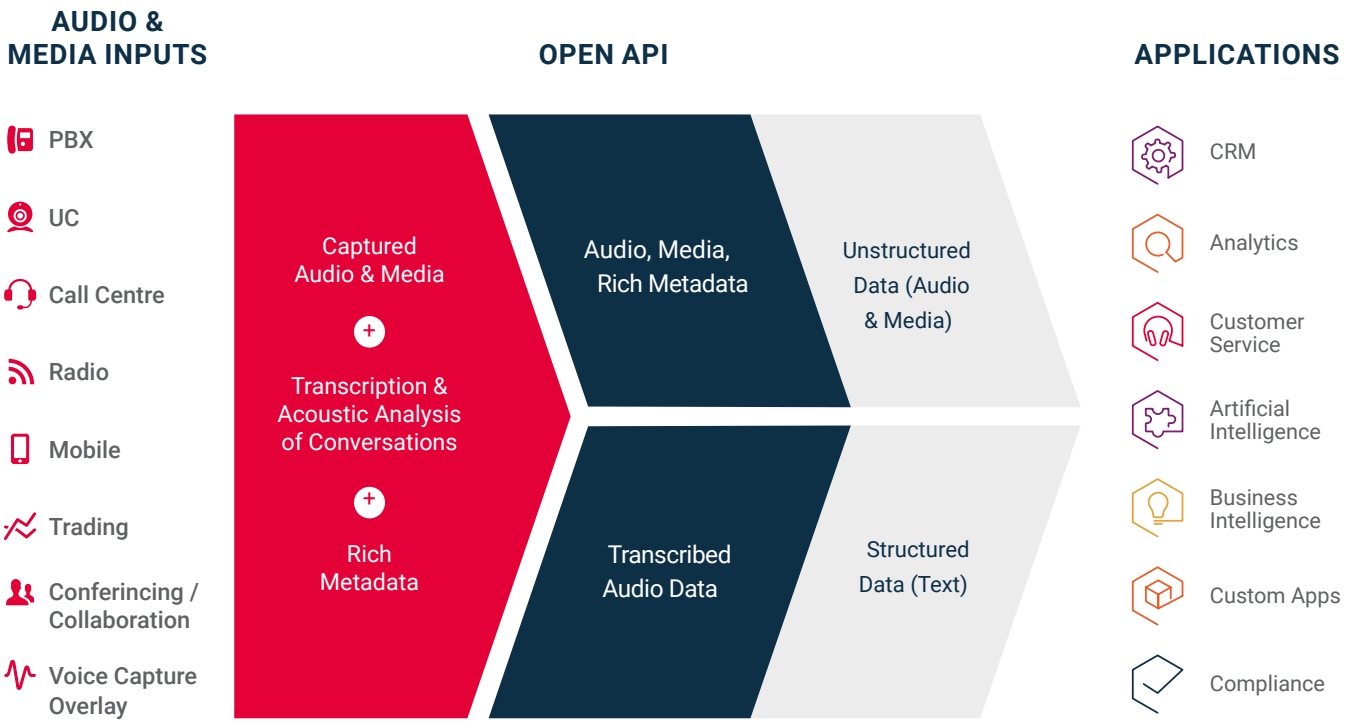
GETTING THE MOST OUT OF YOUR VOICE DATA

To get the most out of metadata, organisations should look for a vendor that provides free access, allowing the data to be filtered into the tools and applications of their choice via open APIs, whether through an extensive ecosystem of leading AI and ML vendors or their own in-house applications.

However, today many traditional vendors are still dictating how and when organisations can use captured voice, often providing compressed and low-quality audio recordings that result in poor transcriptions and sub-par analytics. Many incumbent recording vendors are also charging up to seven figures to export call recordings, often only in batch and not in real-time, preventing the ability to obtain and take advantage of the timely and actionable insights it can produce.

Furthermore, some industries have to retain call recordings for a number of years depending on the regulation in that particular sector, the insurance industry for one. So, when investing in a new call recording system, always ask how you can easily access recordings from older legacy systems alongside more recent recordings should the regulator demand it.

This will become increasingly important as the capture and real-time routing of data with context becomes table stakes in many leading enterprise operations.



CONCLUSION: NOW IS THE TIME TO RE-EVALUATE YOUR RECORDING SYSTEMS

Red Box has always believed that organisations should have secure access to their own call data and an open API approach has always been a core part of our offering. As a result of this openness, we work closely with leading voice analytics software vendors looking to help customers capitalise on this rich stream of data, so we see the real benefits that increasingly sophisticated voice analytics can now offer. However, we are also being approached by a number of software vendors who are looking for help with legacy recording systems that do not work effectively with their customer's investment in AI voice analytics software.

We are seeing that the practices that have been established around recording over the last decades have been entirely based on the assumption that recordings will be listened to by the human ear. Not surprisingly, that has led to a focus on reducing storage costs and using Codecs that often impact voice quality but not to a level that impacts human understanding when listening to individual calls.

Advanced voice analytics based on speech-to-text transcription forces us to re-evaluate that assumption. With so much rich and actionable information held in call recordings, technology can now enable us to mine all this data and gain real insights that can enhance a company's performance. This could be through improved customer experience, process automation, sales performance optimisation or the ability to comply with ever demanding compliance legislations.

IT'S TIME TO RETHINK YOUR RECORDING

But this can only happen if the data foundations are in place to enable software to analyse accurate representations of these conversations in the timescales that make a difference. As we have outlined in this paper, a key dependency is the quality of the recording which may not be possible with your current recording set up.

YOUR CHECKLIST FOR MAXIMISING THE VALUE OF AI

Do you have the right foundations to maximise the benefits of AI and transcription of call data?

If not, you will be limiting the impact of your AI investment!

Uncompressed files

Record in stereo

Real-time data access

Freedom to use any application

ConVersa by Red Box

Bradmore Business Park,
Loughborough Road, Bradmore,
Nottingham NG11 6QA
enquiries@redboxvoice.com
(+44) 115 937 7100 | redboxvoice.com