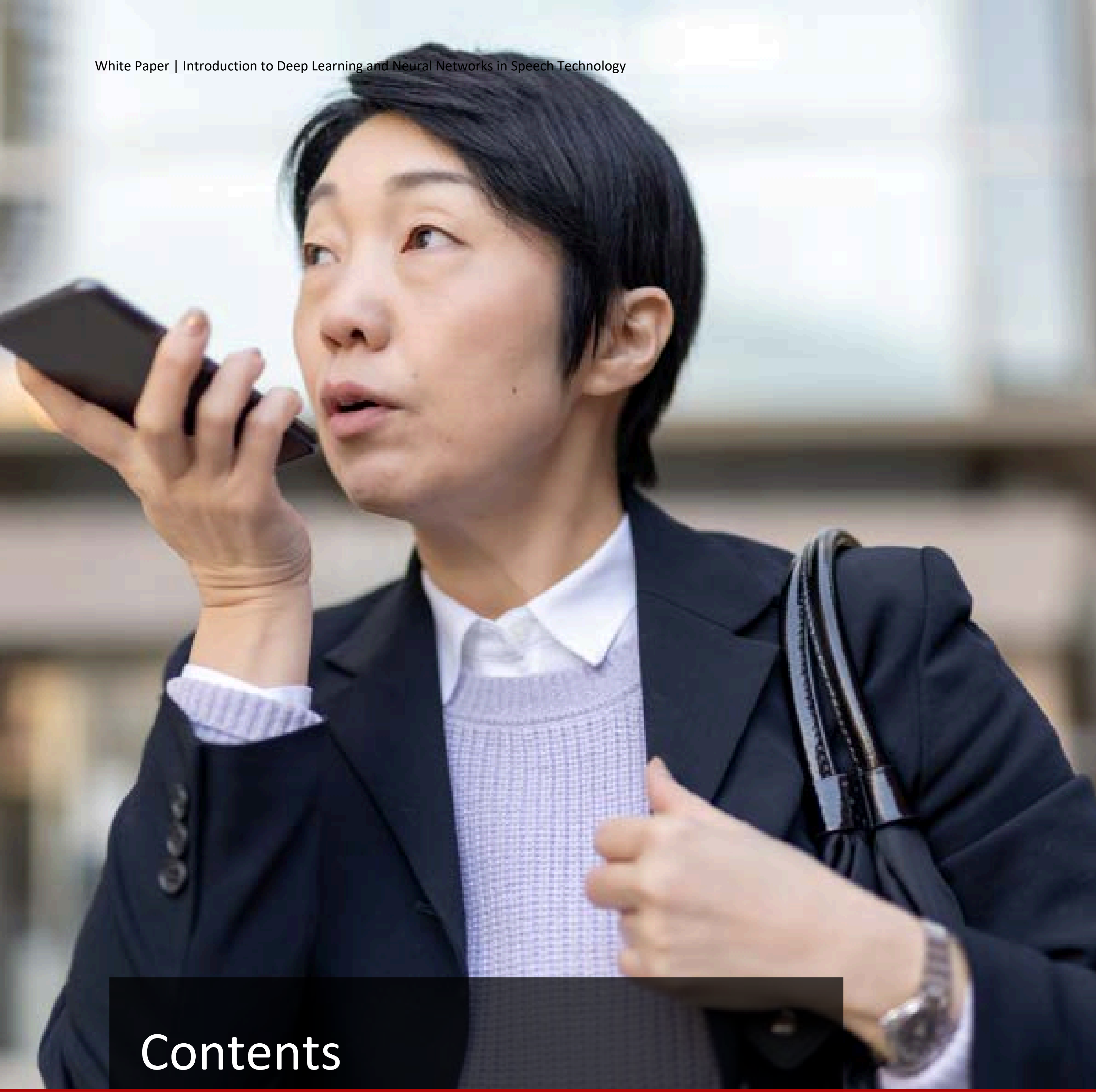
RECOGNOSCO

WHITE PAPER

INTRODUCTION TO DEEP LEARNING AND NEURAL NETWORKS IN SPEECH TECHNOLOGY

An overview of what deep learning is and the benefits to advancing speech technology.

www.recognosco.com



Contents

01

Introduction

Speech Recognition: Bell Laboratories led the way with AUDREY in 1952.

In this digital era, speech recognition technology has evolved as an extremely powerful tool, transforming the way we interact with computers, smartphones, and other devices. In the past, speech technology was based on statistics and probability through machine learning. Today, speech technology has evolved radically with the use of deep learning and neural networks which mimic the workings and modelling of the human brain and how we interpret sounds.

Powered by deep learning and neural networks, speech recognition systems have made remarkable advancements in accurately transcribing human speech into written text. In this white paper, we will explore the many benefits of using speech recognition based on deep learning and neural networks technology, delve into how deep learning models are created, provide insight into future trends and innovations and provide considerations for speech technology partner selection.

In the early 2000s, speech recognition technology began gaining traction, but the accuracy rates were relatively low, and widespread adoption was limited.

1994

The healthcare industry first started implementing speech-recognition systems.

(Source: ResearchGate)

$23.70b

Voice And Speech Recognition Market size value in 2024.

(Source: Grand View Research)

02

The basics: what are neural networks and what is deep learning technology

What are neural networks?

Neural networks are mimicking the most complex and sophisticated object, the human brain. According to a 2009 study, the human brain is made up of roughly 86 billion neurons (pubmed.ncbi.nlm.nih.gov, 2023). A neuron is the simplest unit of any neural network. This is where information processing takes place.

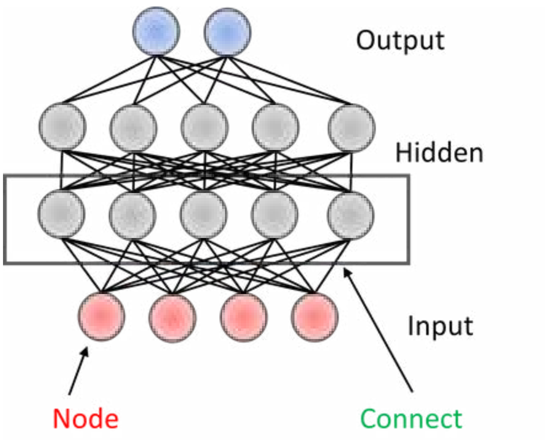
Each neuron combines multiple input signals to create an output signal. These are combined with multiple layers, which each solve a specific task. The first layer – can be compared to the senses of humans and can recognise basic meaning.

In subsequent hidden layers, findings are combined over and over again until a result is available in the output layer. Each layer is optimised with large amounts of data through "trial and error" (working through possibilities to create a conclusion) to detect similarities and create meaning.

Neural networks create a map of input vectors to output vectors. In this way, they can be used for any type of classification of real world data (images, text, sound, time series) which all can be represented as vectors or sequences of vectors.

Figure 1. demonstrates the Deep Neural Networks (DNNs) which include the input layer, the hidden layers and the output layer. It is worth noting that there are many(!) hidden layers – not only the two pictured here as an example.

Figure 1.



In speech technology, the input for the input layer consists of the sounds to be interpreted.

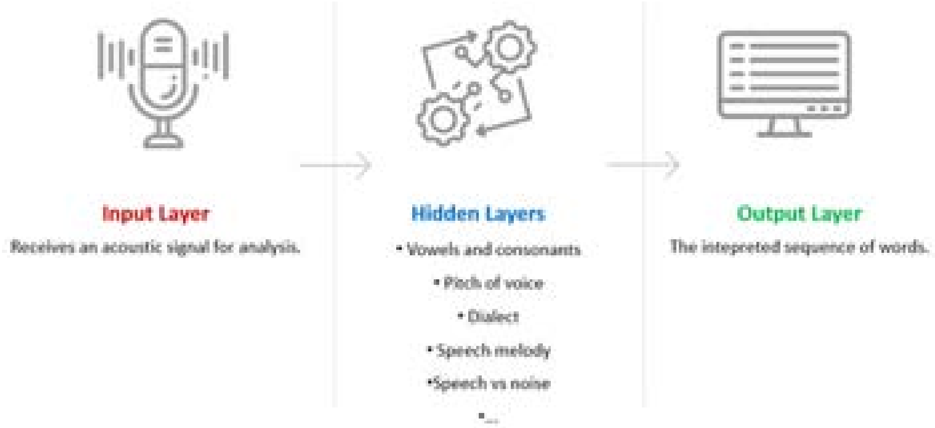
Figure 2. demonstrates an audio file created during a medical dictation. After preprocessing, such data is sent to the input layer of the neural network. This is an example of the pre-processed input for the 'input layer':

Figure 2.

The hidden layers each take the role of depicting vowels and consonants to define words by determining, for example factors including pitch of voice, dialect, speech melody and speech vs. noise. These hidden layers will then come together to determine the interpreted sequence of words with the highest overall similarities. The output layer then presents the defined dictated words. For example, “Patient presented today with right shoulder pain and stiffness.”

Figure 3. is an illustration of the layers within the network.

Figure 3.



Deep Lerning

Deep learning is a type of machine learning inspired by the structure of the human brain. The term "deep" in Deep Learning refers to the large number of layers in the artificial deep neural network. Deep Neural Networks are made possible by the increase of computational power and the availability of large amounts of training data.

Neural networks are at the core of speech technology, allowing us to build highly accurate and adaptable systems for speech recognition, synthesis, and understanding.

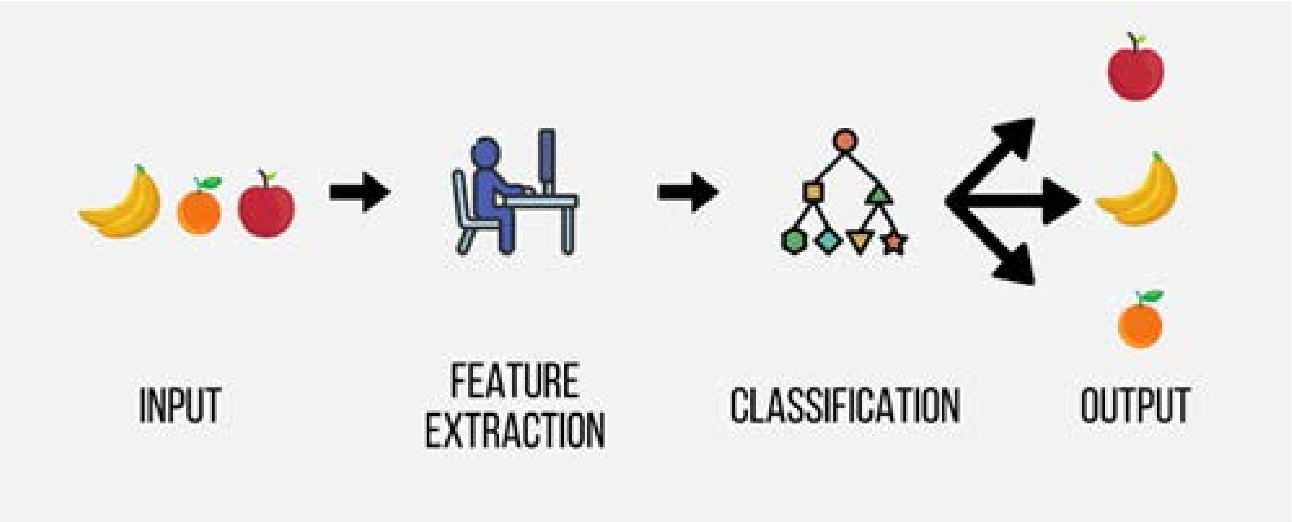
03

Machine learning based speech recognition

Traditionally, speech technology and recognition has been based on machine learning algorithms which use pre-engineered features to develop predictions. For example, based on the input of sounds transmitted, rules, learning and probability create an output sequence of interpreted words and sentences. This is then trained and developed over time to build accuracy.

In Figure 4. from Advancing Analytics, machine learning is demonstrated with fruit sorting. The feature extraction layer is programmed with data to help computers to understand what each fruit looks like and what, for example, constitutes an apple (size, colours, features). This information trains the algorithm to help with the classification of fruits accordingly (based on this data).

Figure 4.



04

The need for Deep Learning-based speech recognition

Deep learning is the next huge step from machine learning. The advancements are like comparing night and day. Deep learning is a subdivision of machine learning that mimics how the human brain operates. Deep learning requires thousands of hours of data to effectively teach computers to do what humans naturally do, learn by example and experience.

In Figure 5. the deep learning model has processed thousands of images of fruits and thus builds a pattern to identify fruits. The 'input' images will be processed through multiple layers of neurons for the 'feature extraction and classification'. Each layer in turn will define the characteristics of the object/word or in this case fruit.

The key differentiator with DNNs is that the output is the total sum of everything – it is not an output based on limited statistics alone. DNNs have the ability to go far beyond, fill gaps and use AI to provide a far more comprehensive and complete picture. For example, in instances where there may be fewer common pronunciations or variations, the network will work through thousands of possibilities to arrive at the most logical output (considering every input detail available).

Figure 5.

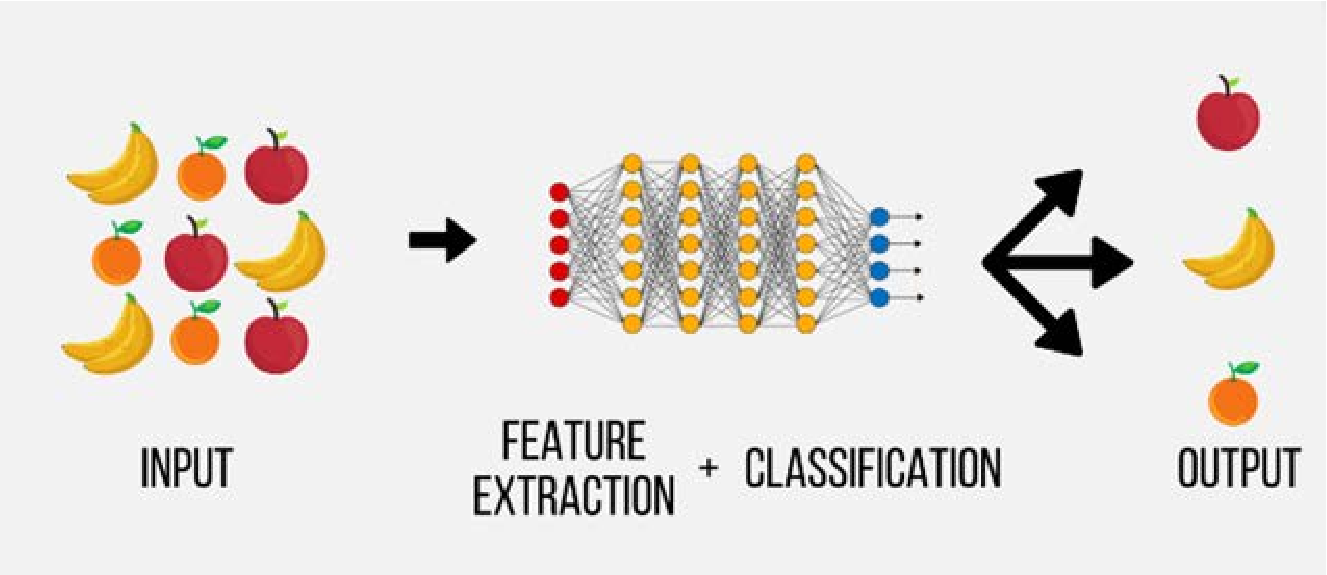


Fig. 4 and 5. Understanding the difference between AI ML and DL using an incredibly simple example Figure on Advancing Analytics. Available from: https://www.advancinganalytics.co.uk/blog/2021/12/15/understanding-the-difference-between-ai-ml-and-dl-using-an-incredibly-simple-example [accessed 1 Oct, 2023]

05

Benefits of deep learning in speech recognition

Since 2010, speech recognition has become commonplace and a part of our daily lives. With the introduction of voice assistants like Siri and Alexa, for example, but also for help to perform voice search queries and send text messages. A recent statistic on usage quoted by Serp Watch (serpwatch.io, 2024), suggests that there are around 4.2 billion digital voice assistants worldwide.

Not only has adoption increased drastically, but with the introduction of neural networks and deep learning, the fundamentals of speech technologies have evolved considerably. Meaning more intelligent, accurate and reliable technology making substantial differences to reporting quality and efficiency.

Prior to the introduction of DNNs, previous machine learning based methods utilised algorithms and probabilities alone. Today's technology is far more comprehensive and all-encompassing, utilising Deep Neural Networks (DNNs) which mimic the workings and modelling of the human brain.

Powered by deep learning and neural networks, today's speech recognition technologies have made remarkable advancements in accurately transcribing speech into written text.

Within this section of the white paper, we will explore some of the benefits of using speech recognition based on deep learning and neural networks technology and delve into the difference this AI-powered technology makes to end users.

Some of the key benefits of utilising a deep learning and neural network-based speech recognition include:

- **Higher Accuracy**
- **Robustness to Environments**
- **Continuous Learning.**

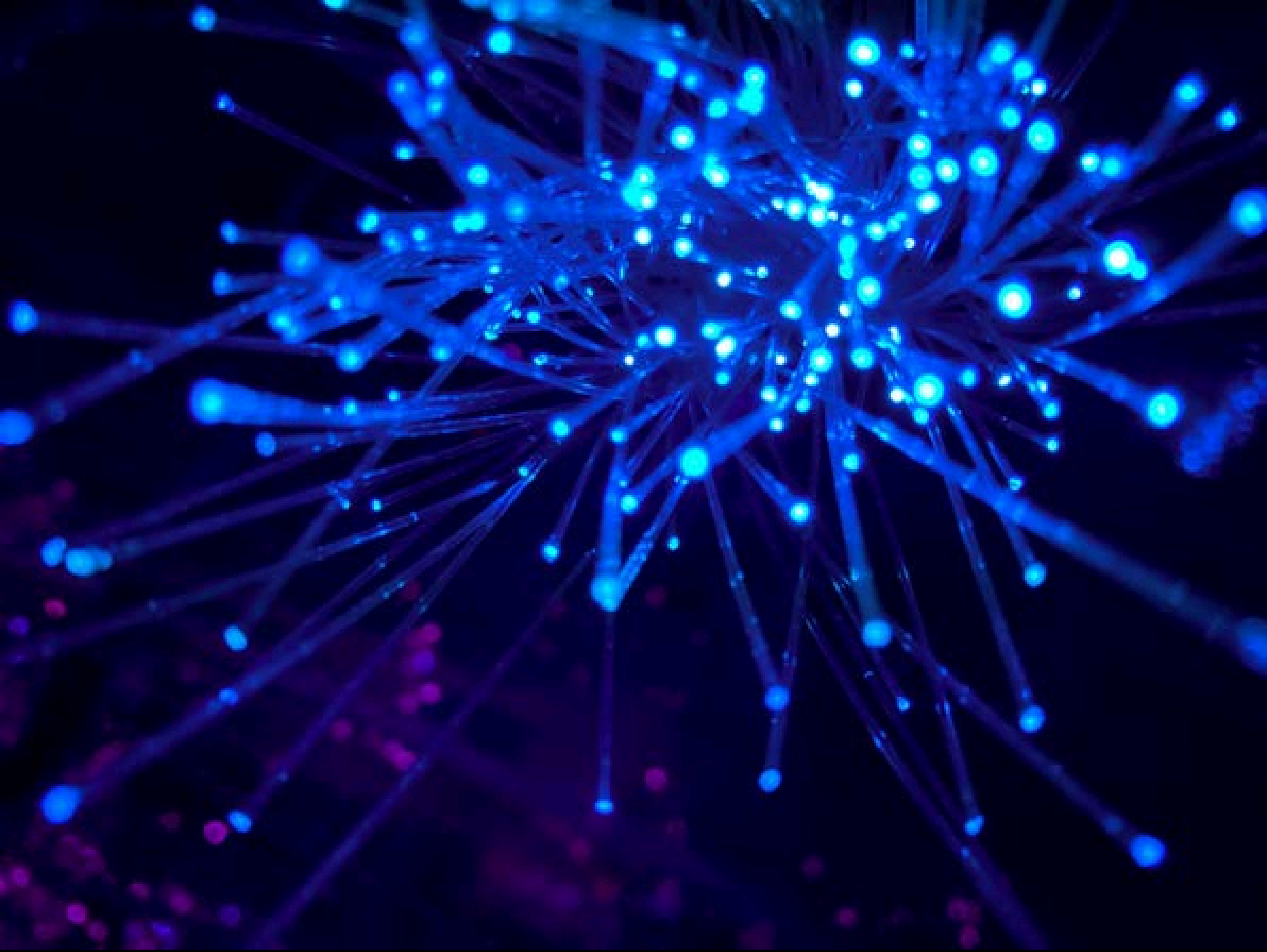
Higher Accuracy

Deep learning techniques, particularly deep neural networks (DNNs), have significantly improved the accuracy of speech recognition outputs. Even with the most complicated terminology and strongest of accents. DNNs can capture complex patterns and features in audio data, resulting in more precise transcription. There is no longer a requirement for voice profiles to be trained to the individual user. Each time a person speaks there will always be differences with pitch, tone, volume etc depending on varying factors including things like the environment, whether the individual has a cold or whether you're dictating fresh in a morning or tired late afternoon.

With the introduction of DNNs, acoustic adaptation happens constantly throughout the duration of a dictation, each time a new report is created. The key differentiator with DNNs is that the output is the total sum of everything – it is not an output based on limited probabilities alone.

Training deep models on vast amounts of data allows the model to use many previous words for the prediction of the next word, which was not possible with classical machine learning. Furthermore the model also continuously extracts a representation of the speaker from the audio input that is then used in the recognition.

For example, in instances where there may be fewer common pronunciations or variations, the network will work through thousands of possibilities to arrive at the most logical output (considering every input detail available). There is no longer a need for the training of voice profiles as the understanding within deep learning and DNNs is far superior. Accuracy levels are of an extremely high standard.



Robustness to Environments and Variations in Speech

Deep learning models can handle variations in speech, such as speakers, accents, dialects, speech impediments, and background noise, more effectively than traditional systems.

With deep learning and DNNs, conclusions are not based on simple probability alone. Neural networks combine the knowledge and understanding of all data provided when considering the output. An example being the English ‘th’ sound. In 10 million cases, this word would be recognised as ‘the’ but this probability doesn’t account for non-native English speakers with accents where this sound may be interpreted as an ‘f’. Modelling accents is very difficult when working on probabilities and averages. Neural networks utilise all information available to consider every scenario to arrive at an output conclusion.

Similarly, when considering background noises when dictating. No two recordings are ever to have exactly the same background noises. There will always be some variations – be that a fan in a busy pathology lab, background conversations, phone calls, animal noises, traffic and more. With the introduction of DNNs, recognition is far greater and robust to background noise.

To achieve robustness in a certain dimension, the training data is augmented by data that varies in that dimension. This can be variations in noise or speakers (Kerle, 2023).

Continuous Learning

It doesn’t just end here. Deep learning models can be continually trained with wider datasets, allowing them to continue to learn, expand and grow.
The benefits that the introduction of DNNs in speech technology bring are significant. The technology evolution benefits clinicians, lawyers, and business professionals across the globe, enabling increased productivity, improved quality and efficiency gains.

06

Training deep learning models for speech recognition

To benefit from the highest levels of transcription, the training of deep learning models is a long and intensive process. This process requires a very large dataset as well as intensive testing. Expert linguists work tirelessly with language models to make sure they are optimised with the highest levels of precision. But what actually goes into the training of deep learning models and what is the process for creating language topics?

These are the steps for training deep learning models:

Data Collection and Preprocessing

Perhaps the most important differentiation when developing deep learning models to previous machine learning based models is the vast dataset that is used in training. Speech samples are gathered that are in the appropriate target language or languages and within the context of your content creation. For example, English UK language and data that is taken from a pathology report. Over time the language we use changes (think of newly discovered illnesses) and words used for reporting may also change. It's important that this context is appropriate and as current as possible – for the best possible results. Data is gathered from a variety of sources ranging from lexica, journals and anonymised reports.

Sophisticated cleaning methods are used to align the textual and audio data prior to training to be able to deal with noise in the textual data.

Training and Optimisation

The procedure of training and optimising the deep learning language models is developed and continually improved by our technologists. This requires much refinement and testing of the results produced. This is the longest and most important part of the process and has to be at an extremely high level to be considered acceptable for quality and assurance.

Evaluation and Tuning

The final part of the training of deep learning models is the evaluation and fine tuning. To ensure we have the highest levels of accuracy, data is constantly revisited, tested, evaluated and regulated. The performance of the trained model is tested and validated measuring things like error rates, precision and recall or other widely used parameters.

Once fully optimised this is rolled out into production environments and performance is continuously monitored. This involves assessing everyday occurrences including changes in audio quality, background noise and different accents.

Challenges

The availability, quantity, cleanliness and quality of the data provided can have a significant impact on the performance of the technology and ultimately the work required to build models to provide the best possible output.

Training deep learning models for speech recognition is computationally intensive, requiring GPUs or similar for faster processing.

Successfully creating and training a speech recognition model is an art. It requires good quality data and vast amounts of it, testing and intensive training, and linguistic and domain-specific knowledge to achieve the best possible results and performance.

```
enu)"),d=b.data("target");if(d)||(d=b.attr(
a.Event("hide.bs.tab",{relatedTarget:b[0]}),g=a.Event("
ented()){var h=a(d);this.activate(b.closest("li"),c),thi
type:"shown.bs.tab",relatedTarget:e[0]})})}}},c.prototyp
ve").removeClass("active").end().find('[data-toggle="ta
ded",!0),h?(b[0].offsetWidth,b.addClass("in")):b.remove
[data-toggle="tab"]').attr("aria-expanded",!0),e&&e()}
.find("> .fade").length);g.length&&h?g.one("bsTransitio
fn.tab;a.fn.tab=b,a.fn.tab.Constructor=c,a.fn.tab.noCo
a(document).on("click.bs.tab.data-api",'[data-toggle="
";function b(b){return this.each(function(){var d=a(th
```

07 Future Trends and Innovations

Things have already progressed significantly with the introduction of AI and DNNs. But what more can we expect in the future?

Technology continues to evolve at a great pace. There will always be new solutions for problems. Data continues to enable us to make great leaps with progression and the introduction of Chat GPT poses vast further opportunities.

Speech technology future innovations.

Multi-language models – one model for all languages. Rather than needing specific models for each individual language, the future may well have the potential to transcribe across multiple languages

Multi-topic models – to gather precision in highly specific scenarios with complex terminology, for example, within rheumatology, previously models had to be created exclusively for that area. Multi-topic models ensure highest levels of accuracy without the need for specific sub-models.

Learning from corrections – continuous learning is available with the introduction of more data and with further training but in the future, we will see models learning as they go from corrections.

08
Speech Tech Partner Selection

Selecting the right technology partner

Choosing the right technology partner for you is crucial. When selecting a partnership, the following factors should be considered:

Expertise

Assess the partner's expertise and experience. Look for a partner who has a strong, evidenced track record of delivering similar successful projects. Consider their technical skills, industry knowledge and the depth of their experience.

Communications and Support

Effective communication and collaboration are critical to successful partnerships. Evaluate service level agreement response times, and ongoing support capabilities. Get a sense of collaboration skills: assessing helpfulness and responsiveness.

Trust and Reliability

Evaluate the partner's commitment to quality and their ability to deliver reliable solutions. Consider their approach to quality assurance, testing processes, and their ability to meet deadlines.

Scalability to flex

Look to understand the partner's ability to scale up or down. Recognosco provides their partners with the option to deploy their speech technology platform on-premises or in Cloud environments. This flexibility is not always available and an important factor for consideration when selecting your technology partner. Does the technology partner offer flexibility of deployment with options for on-premise, in a Cloud environment, or even a combination of both? Assess their ability to adapt to changing project needs and requirements.

Look for a partner that can help you handle future growth and change.

Price and Value

A decision shouldn't be made solely on cost alone but it's an important consideration. Evaluable 'value' in terms of expertise, quality, support and long-term benefits. Consider pricing structure and ensure it aligns with your budget and expected return on investment.

Quality

Get a sense of the partner's commitment to quality and their ability to deliver reliable solutions.

09 Conclusion

Since its inception in 1952, speech recognition technology has progressed considerably. Today's deep learning and neural network-based speech technology, utilising artificial intelligence, is far more sophisticated mimicking the human brain. Deep learning has revolutionised speech technology by significantly improving accuracy and quality.

There are many benefits to utilising a deep learning and neural network-based speech recognition SDK including: higher accuracy, robustness to environments and variations in speech and continuous learning.

As technology advances rapidly, new solutions emerge to address evolving challenges. The availability of data empowers us to make significant strides in progress, and the introduction of ChatGPT opens doors to even greater opportunities for innovation and advancement.

In conclusion, deep learning and neural networks have propelled speech technology to new heights, offering unprecedented accuracy, adaptability, and scalability. By understanding the fundamentals of deep learning and harnessing its potential, software providers can leverage speech technology to enhance user experiences, drive efficiency, and unlock new opportunities.

Interested in learning more about speech-enabling your software? Request more information at: recognosco.com or email: contact@recognosco.net

SCAN ME

Citations and References

- An Analysis of the Implementation and Impact of Speech-Recognition Technology in the Healthcare Sector, ResearchGate: https://www.researchgate.net/publication/5782672_An_Analysis_of_the_Implementation_and_Impact_of_Speech-Recognition_Technology_in_the_Healthcare_Sector [accessed 2 July, 2024)
- Voice And Speech Recognition Market Size, Share & Trends Analysis Report By Function, Grand View Research: https://www.grandviewresearch.com/industry-analysis/voice-recognition-market [accessed 2 July, 2024]
- Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain - PubMed, www.nih.gov [accessed 8 Sep, 2023]
- Human Brain - Structure, Parts, Location, Working, Functions, and FAQs (geeksforgeeks.org)
- Difference between a Neural Network and a Deep Learning System - GeeksforGeeks: https://www.advancinganalytics.co.uk/blog/2021/12/15/understanding-the-difference-between-ai-ml-and-dl-using-an-incredibly-simple-example [accessed 8 Sep, 2023]
- Using Deep Learning to Localize Gravitational Wave Sources - Scientific Figure on ResearchGate. Available from: https://www.researchgate.net/figure/An-Artificial-Neural-Network-ANN-with-two-hidden-layers-and-six-nodes-in-each-hidden_fig1_335855384 [accessed 8 Sep, 2023]
- Lisa Kerle, Michael Pucher, Barbara Schuppler, Speaker Interpolation based Data Augmentation for Automatic Speech Recognition. 20th International Congress of Phonetic Sciences (ICPhS), Prague, Czech Republic., 2023.

RECOGNOSCO

About Recognosco

Providing software partners across the globe with the latest AI-powered speech recognition. Our innovative speech SDK leverages artificial intelligence, utilising Neural Networks and Deep Learning, providing the most advanced speech recognition. Enabling doctors, nurses, lawyers, transcriptionists and agents to produce comprehensive, high-quality documentation - quickly and efficiently.

Based in Vienna, Austria our team has many years of experience in the development, support and distribution of speech recognition technologies for the global medical and legal markets.

Learn more at: recognosco.com
or contact marketing@recognosco.net

Our Partners