altviz

An Introduction to Automatic Speech Recognition (ASR)

Automatic Speech Recognition (ASR)

Automatic Speech Recognition (ASR) is a branch of Artificial Intelligence (AI) that is concerned with converting audio data into text. ASR has made great progress in recent years and has now reached the point where it is usable in industry. If your business generates large amounts of speech data (e.g. call centres), then you should be considering how you can use this data to improve your operations.

This paper looks at the ASR technology, describes how it works and identifies some applications of the technology. It closes with questions that you should be aware of when assessing the technology.

ASR works by taking an audio file and converting it into a sequence of words. This is a complex task, because the audio is stored as a digitisation of an analog signal. Figure 1 show the digitisation of a sound file, i.e. a series of irregular wave patterns.

FIGURE 1 How sound is recorded on a file

If we look closely at the wave pattern, it is not even obvious where the word breaks occur in the data. There are other issues regarding the tempo, volume and frequency range of individual speakers. The ASR solution must be able to deal with people who speak at different speeds, loudness and tones. To solve these challenges there are two approaches in ASR: the traditional approach and the deep learning approach.

Traditional ASR approach

In traditional ASR, the problem is broken down into specific tasks and chained together as a pipeline:

1. Create features from the sound file: this may require windowing (a form of filtering and aggregation) and performing transformations on the windows, such as fourier transformations and connectionist temporal classification (CTC).
2. Apply an acoustic model to match the phonemes*.
3. Apply a language model that uses probability distributions to predict words from the phonemes and then the sequences of words from the phonemes.

The following diagram shows this method.

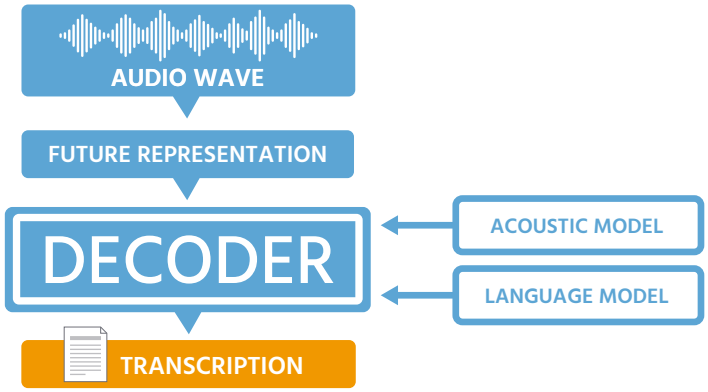

FIGURE 2 Traditional ASR pipeline

ASR toolkits such as Kaldi, CMU Sphinx and HTK all work in this fashion. The advantages of breaking the problem into this form of pipeline is that you can work on each part independently to improve the system. However in practice, this is also a disadvantage as overall the process can be brittle and requires specialised researchers.

Deep learning approach

The second approach to ASR uses deep learning, where the goal is to replace the intermediate steps with one algorithm. The deep learning approach has achieved state of the art results in speech transcription tasks and is replacing the traditional methods used in ASR. It is also simpler because there are fewer steps involved and does not require as much expertise. You will note in the image below that there are two inputs: one is the deep learning toolkit and the second is the audio file. This is the implementation of the ASR module in a deep learning language such as TensorFlow. Some ASR deep learning toolkits include DeepSpeech, PyTorch-Kaldi and CNTK.

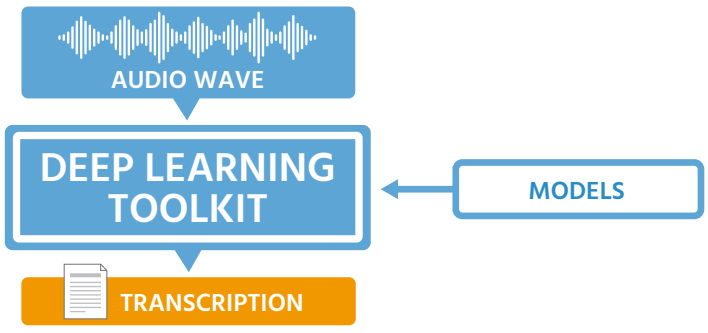

FIGURE 3 Deep learning ASR pipeline

Challenges with these approaches & a new solution

Regardless of whether you select a traditional ASR toolkit like Kaldi, or a deep learning ASR toolkit like DeepSpeech, you will probably need a data scientist to work on the project to get acceptable results. An alternative approach is to use a cloud-based approach to transcribe the data. All of the main cloud providers offer a speech-to-text transcription service. While these provide reasonably accurate results, and are a good way to get started in a project, they will not get the best accuracy compared to a solution that has been optimised for a specific use-case. This is because the machine learning models used by the cloud providers have been trained on generic data rather than domain specific language. You can usually get better results by employing a data scientist to adopt an existing toolkit and optimising it for a particular use-case.

How to measure success

The most recognised method to evaluate ASR is using word error rate (WER), which is the percentage of the number of words correctly recognised**.

For academic tasks on specific datasets, a WER rate of 5% is possible, but for real-life applications, a WER of 10-20% is considered acceptable. This is because the ASR models are trained on historical datasets, which may not reflect modern voice data. Another issue with some models is the inability to handle regional accents, as they may only be trained on voices from the same region (e.g. all native U.S. speakers). This is a particular problem with some of the cloud ASR services.

Post-ASR techniques

Once you have the speech converted to text (known as transcription), the next step is to use this text. This process is known as Natural Language Processing (NLP) and there are various techniques from NLP that you can apply to the text.

Some of these techniques include:

- **NAMED ENTITY EXTRACTION** This involves identifying people, places and organisations.
- **CLASSIFICATION** This involves assigning a category to the audio file.
- **PHRASE MATCHING** This involves looking for a particular phrase within the audio, e.g. statements such as "*your call may be recorded for quality control and training purposes*".
- **NATURAL LANGUAGE UNDERSTANDING** This is analyzing the meaning of text, e.g. taking the named entities and looking at the relations between them.

It is important to realise that if you use a cloud service for transcription, then you probably need to develop these post-ASR techniques. This will require specialised skill sets from a data scientist trained in handling and manipulating text data.

How to evaluate different ASR solutions

When evaluating ASR solutions, here are some questions you should ask:

- Does the software use cloud services under the hood? These cloud service models are generic and not optimised for specific use-cases. A key question is how well it recognises domain specific words and terminology. For example the word "gardai" (Irish for police officer) in motor insurance would probably not be recognised by a cloud transcription service.
- Can the system be deployed on-site and work without connecting to 3rd party services?
- Can the model be extended to cater for new words that were not in the vocabulary used to train the model?
- What is the end result? Simply transcribing the speech to text does not add any value.
- How does the process work with less than perfect transcription?

A practical example of ASR

At Altviz, we built an ASR solution for a customer that monitors at the First Notice of Loss (FNOL) for insurance claims by analysing calls that report motor accidents. We created a domain specific language model and wrote a custom data pipeline that converts the call recording speech to text, extracts key information from the text, merges it with data from CRM systems and classified into type of incident and severity.

This can be used for audit and compliance purposes (e.g. checking that calls were correctly tagged by the call operator). It can also be used to ensure that the call was handled correctly and all the relevant data was selected.

By focusing on a specific use-case, we created a state of the art solution for that application that exceeds the performance of a generic or cloud based model. Now that it is operational, the technology can be adapted for any other solutions that would benefit from using speech recognition.

REFERENCES/APPENDIX:

*Phonemes /ˈfəʊniːm/ — any of the perceptually distinct units of sound in a specified language that distinguish one word from another, for example p, b, d, and t in the English words pad, pat, bad, and bat.

** It is actually calculated slightly differently to this, but that is not important here.

For more information on how Intelligent Automation can help your organisation, please contact us

Richie Barter
CEO & Founder
richie@altviz.co

Ken Clanton
Chief Commercial Officer
ken.clanton@altviz.co

Kevin O'Flynn
Business Development Director Ireland
kevin.oflynn@altviz.co

207–209 Southwark Bridge Road London

www.altviz.co