**Automated Speech Recognition System**

ASR as we know it extends back to 1952 when the infamous Bell Labs created “Audrey,” a digit recognizer. Audrey could only transcribe spoken numbers, but a decade later, researchers improved upon Audrey so that it could transcribe rudimentary spoken words like “hello”.

For most of the past fifteen years, ASR has been powered by classical Machine Learning technologies like Hidden Markov Models. Though once the industry standard, accuracy of these classical models had plateaued in recent years, opening the door for new approaches powered by advanced Deep Learning technology that’s also been behind the progress in other fields such as self-driving cars.

In 2014, Baidu published the paper, Deep Speech: Scaling up end-to-end speech recognition. In this paper, the researchers demonstrated the strength of applying Deep Learning research to power state-of-the-art, accurate speech recognition models. The paper kicked off a renaissance in the field of ASR, popularizing the Deep Learning approach and pushing model accuracy past the plateau and closer to human level.

Not only has accuracy skyrocketed, but access to ASR technology has also improved dramatically. Ten years ago, customers would have to engage in lengthy, expensive enterprise speech recognition software contracts to license ASR technology. Today, developers, startup companies, and Fortune 500s have access to state-of-the-art ASR technology via simple APIs like AssemblyAI’s Speech-to-Text API.

Let’s look more closely at these two dominant approaches to ASR.

**How ASR Works**

Today, there are two main approaches to Automatic Speech Recognition: a traditional hybrid approach and an end-to-end Deep Learning approach.

**Traditional Hybrid Approach**

The traditional hybrid approach is the legacy approach to Speech Recognition and has dominated the field for the past fifteen years. Many companies still rely on this traditional hybrid approach simply because it’s the way it has always been done--there is more knowledge around how to build a robust model because of the extensive research and training data available, despite plateaus in accuracy.

Here’s how it works:

**Traditional HMM and GMM systems**

Traditional HMM (Hidden Markov Models) and GMM (Gaussian Mixture Models) require forced aligned data. Force alignment is the process of taking the text transcription of an audio speech segment and determining where in time particular words occur in the speech segment.

As you can see in the above illustration, this approach combines a lexicon model + an acoustic model + a language model to make transcription predictions.

Each step is defined in more detail below:

Lexicon Model

The lexicon model describes how words are pronounced phonetically. You usually need a custom phoneme set for each language, handcrafted by expert phoneticians.

Acoustic Model

The acoustic model (AM), models the acoustic patterns of speech. The job of the acoustic model is to predict which sound or phoneme is being spoken at each speech segment from the forced aligned data. The acoustic model is usually of an HMM or GMM variant.

Language Model

The language model (LM) models the statistics of language. It learns which sequences of words are most likely to be spoken, and its job is to predict which words will follow on from the current words and with what probability.

Decoding

Decoding is a process of utilizing the lexicon, acoustic, and language model to produce a transcript.

Downsides of Using the Traditional Hybrid Approach

Though still widely used, the traditional hybrid approach to Speech Recognition does have a few drawbacks. Lower accuracy, as discussed previously, is the biggest. In addition, each model must be trained independently, making them time and labor intensive. Forced aligned data is also difficult to come by and a significant amount of human labor is needed, making them less accessible. Finally, experts are needed to build a custom phonetic set in order to boost the model’s accuracy.

#End-to-End Deep Learning Approach

An end-to-end Deep Learning approach is a newer way of thinking about ASR, and how we approach ASR here at AssemblyAI.

**How End-to-End Deep Learning Models Work**

With an end-to-end system, you can directly map a sequence of input acoustic features into a sequence of words. The data does not need to be force-aligned. Depending on the architecture, a Deep Learning system can be trained to produce accurate transcripts without a lexicon model and language model, although language models can help produce more accurate results.

CTC, LAS, and RNNT

CTC, LAS, and RNNTs are popular Speech Recognition end-to-end Deep Learning architectures. These systems can be trained to produce super accurate results without needing force aligned data, lexicon models, and language models.

**Advantages of End-to-End Deep Learning Models**

End-to-end Deep Learning models are easier to train and require less human labor than a traditional approach. They are also more accurate than the traditional models being used today.

The Deep Learning research community is actively searching for ways to constantly improve these models using the latest research as well, so there’s no concern of accuracy plateaus any time soon--in fact, we’ll see Deep Learning models reach human level accuracy in the next few years.

#ASR Key Terms and Features

Acoustic Model: The acoustic model takes in audio waveforms and predicts what words are present in the waveform.

Language Model: The language model can be used to help guide and correct the acoustic model's predictions.

Word Error Rate: The industry standard measurement of how accurate an ASR transcription is, as compared to a human transcription.

Speaker Diarization: Answers the question, who spoke when? Also referred to as speaker labels.

Custom Vocabulary: Also referred to as Word Boost, custom vocabulary boosts accuracy for a list of specific keywords or phrases when transcribing an audio file.

Sentiment Analysis: The sentiment, typically positive, negative, or neutral, of specific speech segments in an audio or video file.

See more models specific to AssemblyAI.

#Key Applications of ASR

The immense advances in the field of ASR has seen a correlation of growth in Speech-to-Text APIs. Companies are using ASR technology for Speech-to-Text applications across a diverse range of industries. Some examples include:

Telephony: Call tracking, cloud phone solutions, and contact centers need accurate transcriptions, as well as innovative analytical features like Conversation Intelligence, call analytics, speaker diarization, and more.

Video Platforms: Real-time and asynchronous video captioning are industry standard. Video editing platforms (and video editors alike) also need content categorization and content moderation to improve accessibility and search.

Media Monitoring: Speech-to-Text APIs can help broadcast TV, podcasts, radio, and more quickly and accurately detect brand and other topic mentions for better advertising.

Virtual Meetings: Meeting platforms like Zoom, Google Meet, WebEx, and more need accurate transcriptions and the ability to analyze this content to drive key insights and action.