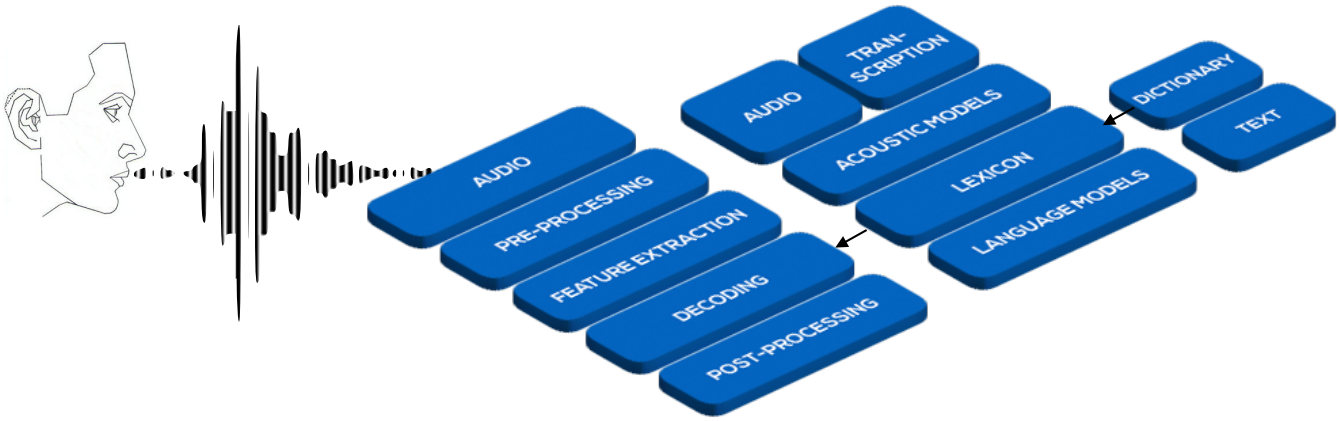

AppTek
ASR WHITE PAPER

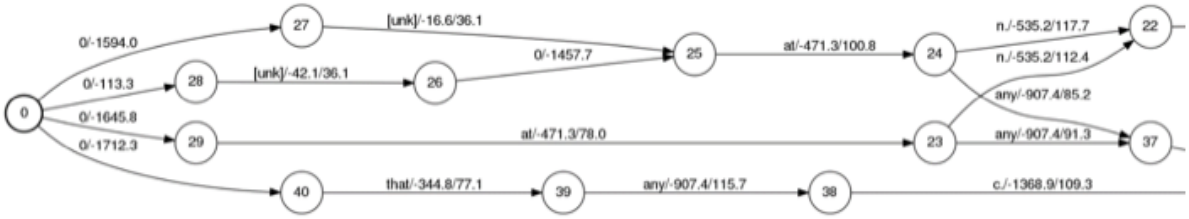
General Speech Recognition Process

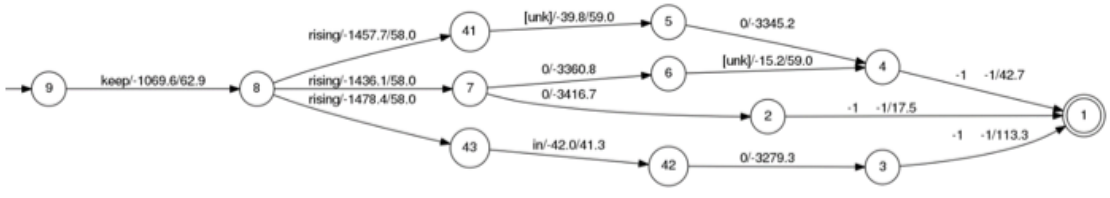
A speech recognition engine includes the following components: acoustic feature extractors, acoustic models, language models, lexicon, decoders, and post-processing (see the figure below). The feature extraction converts waveform into feature vectors; the decoders take the feature vector as input and search for optimum paths using the acoustic models, language models, and the lexicon; and the post-processing analyzes and processes the recognition output. The decoder component might involve combination of different models and/or recognition in a multiple-pass fashion.



Lattice Generation

During recognition, the decoder keeps many hypotheses active simultaneously. The number of hypotheses is defined with a parameter named **beam width**. These hypotheses are interconnected and form a lattice as shown in the figure below.



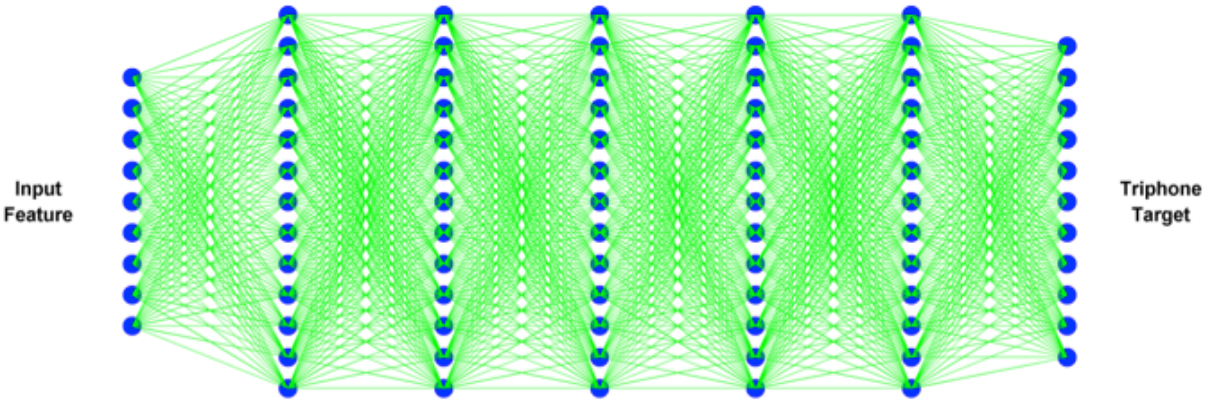


Confidence Scoring

As seen in the above figure, there are alternative paths in the lattice. Confidence scores are computed for each word in lattice indicating how easily each word can be substituted with another candidate. That is, if a hypothesis word (or the portion in the path) in the lattice is very unique, then its confidence score is high, and otherwise low.

Deep Neural Networks Modeling

At AppTek, acoustic models were trained with deep neural networks. Deep neural networks are simulations of human neurons with many layers of processing. In recent years, training a deep neural networks model with thousands of hours of audio is made more practical by leveraging Graphics Processing Unit (GPU) computing technologies.



Keyword Search Using Lattices And Confidence Scores

In our keyword search application, the keywords are searched against the generated lattice during the speech recognition process. The advantage is that although the keyword might not be in the top-best speech recognition result, it could be in the lattice. This will significantly improve the recall rate of finding the keywords. At the same time, confidence score is used to improve the precision of the search: Only words with confidences above certain threshold will be returned. Nevertheless, the confidence threshold will be set to balance the recall and precision rate.