Классический подход.

Классический подход к распознаванию речи представляет собой последовательность следующих действий.

1. На входе у нас есть некая входная речь, то есть график зависимости амплитуды колебаний от времени. Чтобы извлечь полезную информацию из нее нам необходимо как-то преобразовать эти данные.
2. Для этого график разбивается на много маленьких кусочков - фреймов (обычно по 10-30 миллисекунд) и далее происходит выделение признаков (то есть преобразование фрейма в вещественный набор признаков) с помощью теории обработки сигналов. Стандартым методом вычленения признаков является MFCC (Мел-Кепстральные коэффициенты)
3. Далее полученные признаки подаются на вход акустической модели, которая используя, например, скрытые марковские модели (HMM) или глубокие нейронные сети (DNN), получает на выходе распределение вероятностей различных фонем/сенонов на этом участке. Таким образом, каждому кусочку звука будет соотнесено свое распределение вероятностей принадлежности к той или иной фонеме (сенону, звуку)
4. Далее распределение вероятностей подается на вход декодеру, выбирающий наиболее вероятный путь перехода по HCLG-графу, который состоит из 4 частей: модуль на базе акустической модели, модуль контекстной зависимости, модуль произношения и модуль языковой модели.
5. И на последнем этапе происходит перевзвешивание гипотиз и выдача результата.

Проблема такого подхода в том, что акустические модели, модели контекстной зависимости и языковые модели обучаются раздельно.

End2End подход.

End2End подразумевает систему, использующую в качестве обучения критерий, зависимый от конечной метрики, интересующей в данной задаче (word error rate).