Мозг, время, потоковая обработка и пустые состояния в модели CTC.

Чем дольше мы изучаем реальность, тем необычней она нам кажется. Например, по текущим представлениям, мозг обладает следующими свойствами:

мозг – высокопараллельная система, информация в мозгу передаётся с помощью импульсов. Эти принципы, отличающиеся от обычного линейного подхода классической физики, можно учитывать при проектировании речевых систем.

Древние греки знали три типа времени: Хронос – непрерывное линейное время, Циклос –повторяющаяся последовательность событий, и Кайрос – уникальный момент. Со времен Ньютона все свелось к линейному времени, но в середине 20-го века начали рассматривать другие типы времени. Квантовая механика и изобретение компьютеров совершили большой прорыв в подходе ко времени. Здесь можно упомянуть одну из фундаментальных работ Лесли Лэмпорта по логическому времени «Время, часы и порядок событий в распределенной системе». Если мы рассматриваем мозг как высокопараллельную систему, то фактически это означает, что время в мозге действительно нелинейно. Это скорее Циклос, чем Хронос.

Идея импульса неявно присутствует в современных моделях распознавании речи. В процессе создания таких систем мы перешли от HMM с 9-ю состояниями для распознавания цифр TIDIGTS к HMM с 3-я состояниями для систем слитной речи на основе GMM-HMM, а затем к архитектуре CTC с двумя состояниями, которая более точно соответствует функции мозга: есть пустое состояние и состояние генерации импульса, соответствующее звуку. Архитектура CTC эффективна как с точки зрения необходимой памяти, так и скорости вычислений. Есть и недостатки: при переходе в пустое состояние CTC может пропускать слова в шумовых входных данных. Хотя публикаций по CTC архитектуре существует много, теоретические основания часто не упоминаются. Например, если вы читаете оригинальную статью по CTC, у вас может сложится впечатление, что эта архитектура – всего лишь удачное предположение авторов.

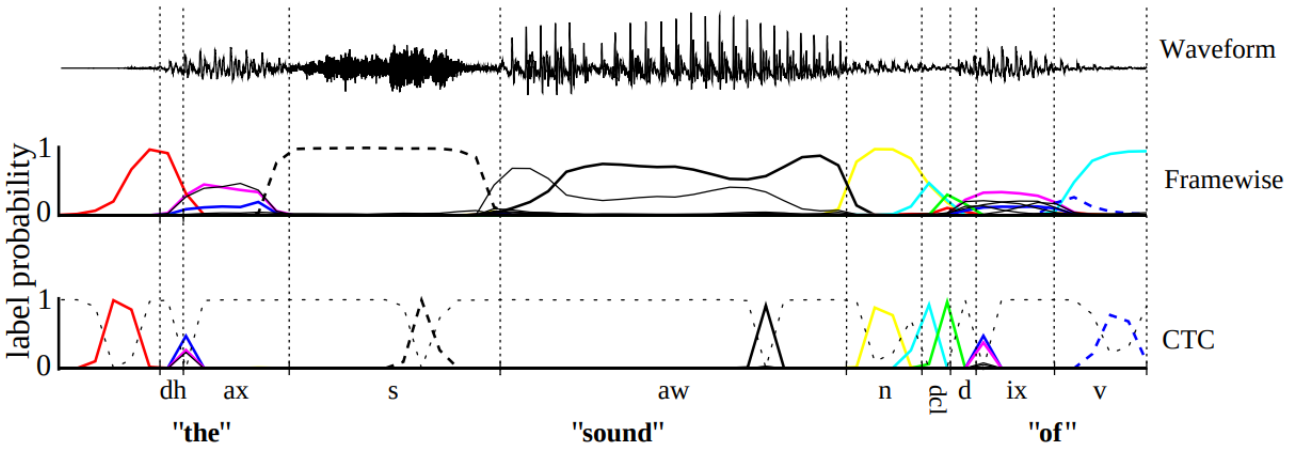

Рис. 1: Модель CTC.

Идея импульса используется не только в распознавании речи, но и в других областях. В архитектуре кодировщика текста системы синтеза речи Glow-TTS тоже есть пустое состояние, которое также используется в VITS1, а затем без особых объяснений отбрасывается в VITS2. Многие задаются вопросом, для чего это нужно. Автор кода утверждает, что никаких оснований нет. На наш взгляд, решение об отказе от пустых состояний в VITS2 выглядит неправильным. Наши эксперименты по обучению моделей VITS2 с пустым состоянием и без него подтверждают это.

Еще одна область использования импульсов – по-прежнему распознавание речи, но с более высокоуровневыми импульсами соответствующими смене контекста, эмоциям, ударениям. Современные языковые модели используют так называемые метки для распознавания языка и специальных событий в тексте. В системе Whisper метки используются для моделирования начала перевода на другой язык и других специальных событий. Метки работают эффективнее, чем кодирование языка записи в виде дополнительных признаков, но мы редко думаем о метках, как об импульсах.

В соответствии с концепцией мозга при проектировании систем распознавания речи нам необходимо пересмотреть подход к потоковой обработке. Мы долгое время были сторонниками быстрого потокового распознавания с низкой задержкой, когда речь обрабатывается маленькими отрезками по 40 миллисекунд с небольшой задержкой для контекста примерно в 500 миллисекунд. Несмотря на все усилия, такие системы не могут достичь необходимой точности распознавания. Процент ошибок в системах с потоковой обработкой речи намного выше, чем непотоковой.

Учитывая параллельность мозга, необходимо обрабатывать входящие речевые данные параллельно, то есть нам не обязательно всегда генерировать ответ с очень низкой задержкой. У нас могут быть определенные ориентиры во времени (например, импульсы), когда мы обновляем результат распознавания с помощью параллельной обработки. Некоторые разработчики, например, в Whisper и Silero уже используют такой подход.

В краткосрочной перспективе триггером может служить технология VAD. В более долгосрочной перспективе мы можем построить триггерную сеть, которая будет запускать параллельную обработку в моменты низкой энтропии или в другие важные моменты, но не слишком часто.

Таким образом, вместо экспериментов с архитектурами, мы можем попытаться разработать архитектуру распознавание речи, основанную на понимании механики мозга. Понимание принципов работы мозга обосновать выбор нейросети и алгоритмов, и двигаться в правильном направлении.

Конечно, есть много исследований на эту тему, но они недостаточно знакомы и понятны речевым инженерам. Например, Enhancement of speech-in-noise comprehension through vibrotactile stimulation at the syllabic rate.