Санкт-Петербургский Государственный Университет
Математико-механический факультет

Кафедра системного программирования

Вельямидов Виталий Альбертович

Хранение и поиск N-грамм для распознавания речи в системе с большим объемом оперативной памяти

Курсовая работа

Научный руководитель:

Санкт-Петербург
2015

Оглавление

Введение

Распознавание речи является одним из важных направлений искусственного интеллекта. Использование такого инструмента уже очень распространено: от приложений на мобильных устройствах, управления ими, до выведения субтитров в прямом видео-эфире.

Однако, повсеместное использование речи делает ее распознавание весьма сложным. Эта сложность возникает из-за сильного различия аудио-сигнала, передающего речь в зависимости от многих факторов. Например, влияет тип действия, которое определяет речь. Это может быть простое общение, чтение книги, выступление и т.д. Также сигнал зависит от темпа и громкости, особенностей произнесения и акцента, человеческие дефекты речи. Усложняет распознавание и зашумленность сигнала, наличие эхо и, так называемый, эффект Ломбарда — небольшое изменение голоса и его громкости при разговоре в шумном окружении. А распознавание детского голоса иногда и вовсе производят с использованием отдельно обученной системы.

Сложно вообразить непосредственно прямое преобразование цифрового сигнала в текст. Поэтому в общем случае система распознавания речи является многокомпонентной. Общую схему можно увидеть на рис. 1.

Первым этапом распознавания является представление входного цифрового сигнала в виде вектора числовых параметров (feature extraction). Этот этап очень важен, т.к. именно с полученными характеристиками будут производиться все дальнейшие действия. Основные методы для реализации этого этапа описаны в статье [1].

Далее на отдельно взятом участке необходимо выделить фонемы - звуковые единицы речи, и составить из них фонетические слова. Абсолютно точно их определить пока что не представляется возможным, поэтому используется вероятностная акустическая модель, которую можно представить в виде $P(O|W)$, где O - параметризованное представление взятого участка, а W - возможные слова в языке. Иногда сюда же добавляют модель произношения, которая по своей сути являет-

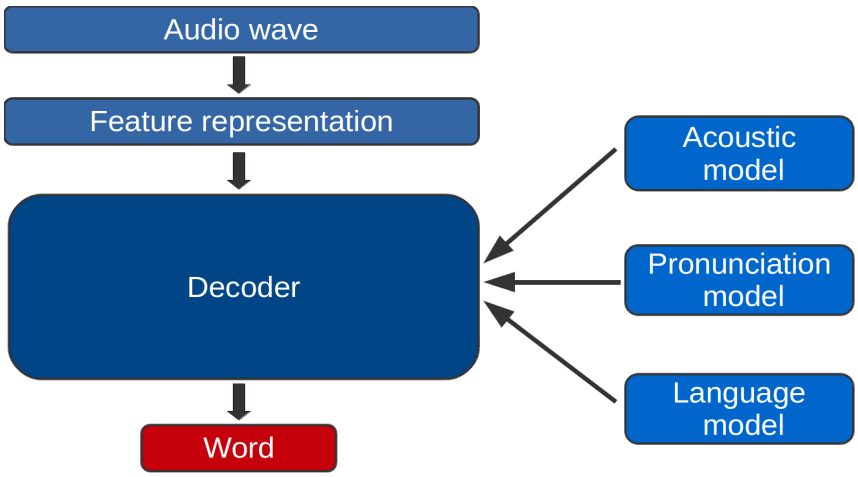

Рис. 1: Общая архитектура системы распознавания речи

ся словарем Q, отображающим каждое слово в возможные варианты его произношения. Тогда совокупность модели произношения с акустической имеет вид $P(O|Q(W))$. Что касается реализации акустической модели — существует два основных подхода: использование скрытой Марковской модели [2] или построение нейронной сети [3].

Если каждый раз выбирать слово, на котором достигается максимальна вероятность в акустической модели, то, скорее всего, текст не будет логически связным. Произнесенное слово может быть очень похоже на другое, которое в данном контексте не уместно, поэтому разрабатывают и используют языковую модель, которая отвечает за возможность появления слова в конкретном контексте. В общем виде ее можно представить как $P(W)$. Именно этот этап является узким местом в производительности системы распознавания речи. Это обуславливается тем, что языковая модель обучается на большом корпусе, и обучаться должна быстро, для обновления языковой базы, а так же, что самое главное, необходим быстрый поиск по этой мощной модели, поскольку даже для распознавания одного предложения требуется выполнить к ней множество запросов. Самым распространенным вариантом такой модели является использование N-грамм - цепочек слов длины N. Тогда для каждого слова определяется вероятность его появления в зависимости от N предыдущих слов [4].

Таким образом, считая, что в процессе распознавания необходимо взять наиболее вероятный вариант для заданного наблюдения O, получить такое слово можно по формуле (1).

$$\underset{W}{\operatorname{argmax}} P(O|Q(W))P(W) \tag{1}$$

К сегодняшнему моменту для сложных расчетов уже набрали популярность многоядерные системы, способные выполнять несколько подзадач одновременно. Также сейчас никого не удивишь вычислительными системами с большим объемом оперативной памяти. Они, например, часто используются для обработки большого количества статистических данных. Зная, что построение языковой модели оперирует с большим корпусом, а поиск по построенной модели имеет достаточно высокий уровень сложности, хочется проверить, как будут вести себя различные реализации языковых модели на подобных многоядерных системах с большим количеством RAM. А также попробовать улучшить одну из них или реализовать принципиально новый алгоритм, адаптированный под такие системы, что и является целями дальнейшей научной работы.

1. Хранение и поиск N-грамм

Одно из основных направлений исследовательской деятельности в области языковых моделей на основе N-грамм заключается в том, чтобы добиться максимально компактного хранения модели, и быстрого поиска N-грамм в ней. Например, это можно наблюдать в одних из последних статях исследователей из университета Цукуба (Япония) [5] и университета Калифорнии [6].

Пока что есть две общие идеи для повышения производительности операций с языковой моделью на системах с большим объемом оперативной памяти:

1. Подобрать такую структуру данных и ее реализацию, чтобы, жертвуя количеством памяти на хранение модели, увеличить скорость выполнения запросов.

2. Проанализировать последовательность выполнения поисковых запросов, общую суть совокупности запросов, и попытаться сгруппировать их и организовать поиск таким образом, что бы за один 'новый' запрос получать больше информации, т.е. ответы на группу 'старых' запросов

За имеющийся пример языковой модели, для сравнения, можно взять уже существующую и широко распространенную систему SRILM, суть и функциональность которой описана в статьях [7] и [8].

Список литературы

[1] K. R. Ghule and R. R. Deshmukh, "Feature extraction techniques for speech recognition: A review," *International Journal of Scientific and Engineering Research*, vol. 6, pp. 258–267, May 2015.

[2] M. Gales and S. Young, "The application of hidden markov models in speech recognition," *Foundations and Trends in Signal Processing*, vol. 1, pp. 195–304, January 2008.

[3] G. Hinton, L. Deng, D. Yu, A. rahman Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. S. G. Dahl, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition," *IEEE Signal Processing Magazine*, vol. 29, pp. 82–97, November 2012.

[4] D. Jurafsky and H. James, *Speech and language processing.* An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition, Alan Apt, 1999.

[5] M. Yasuhara, T. Tanaka, J. Norimatsu, and M. Yamamoto, "An efficient language model using double-array structures," in *EMNLP*, pp. 258–267, 2013.

[6] A. Paus and D. Klein, "Faster and smaller n-gramm language models," in *HLT '11 Proceedings of the 49th Annual Meeting of the Association for Computational*, pp. 258–267, 2011.

[7] A. Stolcke, "Srilm - an extensible language modeling toolkit," in *International Conference on Spoken Language Processing*, pp. 901–904, 2002.

[8] A. Stolcke, J. Zheng, W. Wang, and V. Abrash, "Srilm at sixteen: Update and outlook," in *in Proceedings IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, pp. 5–9, 2011.