1 Статьи

Статьи записаны в хронологическом порядке.

1. A Bilingual Kazakh-Russian System for Automatic Speech Recognition and Synthesis, 2015
2. Continuous speech recognition of kazakh language, 2018
3. Impact of Using a Bilingual Model on Kazakh–Russian Code-Switching Speech Recognition, 2019
4. A Crowdsourced Open-Source Kazakh Speech Corpusand Initial Speech Recognition Baseline, 2020
5. End-to-End Speech Recognition in AgglutinativeLanguages, 2020
6. A Study of Multilingual End-to-End Speech Recognition for Kazakh, Russian, and English, 2021
7. Hybrid end-to-end model for Kazakh speech recognition, 2022

1.1 Статьи, которые показались менее интересными

В этот список я включу статьи под номерами 2, 3, 5.

- В статье 2 приведены только эксперименты со 'старенькой' архитектурой DNN-HMM на Kaldi в различных вариациях. Не используют аугментации, языковые модели. Странным решением показалось перевод казахских букв в английские. Мотивацию не понял. Нет кода, нет датасета.
- В статье 3 так же обучают только DNN-HMM. Авторы сравнивают WER моноязычных моделей, обученных на русском и казахском датасетах, и мультиязычной, обученную на обоих датасетах вместе. Задача минимум выполнена, буст мультиязычной модели на казахском датасете в несколько пунктов WER показан, но само значение WER около 50 сильно смущает... Кода нет, датасетов нет.
- В статье 5 авторы сравнили архитертуры типа MLP, CNN, LSTM, ResNet на казахском датасете, но забыли о трансформерах. Делали эксперименты с LM и MFCC. Ничего новаторского, но это уже достаточное количество экспериментов в сравнении с предыдущими статьями. Из минусов могу отметить отсутствие аугментаций и экспериментов с BPE

1.2 Статьи, которые привлекли внимание

Перед тем как переходить к статьям хочется отметить, на сколько я понимаю, сложность и мотивацию этого проекта. Так как много людей в Казахстане используют в своей речи русские слова, то нам придется использовать некие хитрости, чтобы не слишком часто ошибаться в этих местах. Хитрости могут быть связаны с различными датасетами, а так же с изменёнными словарями токенов. Посмотрим, что это могут быть за хитрости:

1.2.1 Ребята из ISSAI

В этой части я решил объединить статьи 4 и 6, так как они написаны практически одними и теми же людьми из одного института. У них есть гитхабы, первый и второй, соотвественно. Не скажу, что они супер полезные, но они есть. Для обучения используют ESPnet и Kaldi. Авторы используют датасет на ∼400 часов аудио(в настоящее время он около 1200 часов), который надо запрашивать. Я запросил, пока жду ответа. Нашел еще их отдельный датасет для TTS(тоже запросил) (сейчас он включен в этот новый 1200 часовой датасет).
В статье 4 авторы сравнивают DNN-HMM, LSTM и трансформеры(энкодер+декодер).
В качестве инпутов они подают MFCC, делают speed perturbation, and spectral augmentation.
Также скор бустят языковой моделью на буквах. В итоге, как можно догадаться, победил трансформер. Даже попробовали BPE, но они почему-то они не зашли. Из нового, они решили пересечь русский и казахский алфавиты(одинаковые буквы из разных алфавитов сливаются в одну, а разные так и остаются), и в итоге они получили WER порядка 9-10 в лучшем сеттинге. Единственное, не понятно, дает ли совмещение алфавитов какой-то буст по сравнению с один казахским алфавитом, потому что такого эксперимента не приведено.
В статье 6 они пошли дальше, взяли трансформер(энкодер+декодер) и решили обучить мультиязычную модель на объединении 3-х датасетов: казахского, русского и английского. Словарь составили из объединения всех трех алфавитов, в отличии от предыдущего подхода. Я думаю, не получив буста с BPE в прошлый раз, в этот раз они решили не проводить эксперименты с ними. Даже в случае моноязычной модели они смогли выбить WER на 0.5 пункта, чем в прошлой статье. К сожалению, мультиязычной модель показала на казахском датасете такой же скор, как и моноязычная. Стоило попробовать не обучать на английском датасете, потому что, наверное, он только зашумляет предсказания.

1.2.2 A Bilingual Kazakh-Russian System for Automatic Speech Recognition and Synthesis

Это статья под номером 1. Еще в 2015 году авторы, используя DNN-HMM, предложили сначала обучать модель на русском корпусе, а потом дообучать на казахском. Больше здесь не было ничего интересного. Только я бы в этом случае, так как мы делаем ASR для казахской речи, сначала обучался на казахском корпусе, а потом дообучался на русском до тех пор, пока модель показывает хорошее качество.

1.2.3 Hybrid end-to-end model for Kazakh speech recognition

Эта статья под номером 7. Самая новая и актуальная из моего списка.
Авторы предложили некий новый подход к построению предсказаний. Они подают сырую 'вафку' в энкодер, а выходы из него попадают и в CTC, и в какой-то декодер с attention(там не сказано, что именно, но я полагаю, что по аналогии с LAS). Итоговые предсказания считаются как взвешенная сумма предсказаний из CTC и декодера. Такое решение идейно похоже на идею RNN-T, где впервые совместили плюсы CTC и энкодер-декодер архитектуры. Их модифицированная архитектура побила, конечно, CNN, LSTM и Resnet, но скоры были посчитаны на разных по количеству(и скорее всего по сути) примерах в датасетах, а это, кажется, не очень правильно. Плюс к этому всему они не сравнили скоры с трансформерами.

2 Датасеты

Как я уже говорил, я нашел хороший датасет для казахского языка, но пока не могу на него посмотреть. Предполагаю, что он выглядит так как стандартный датасет для ASR - WaV + Текст. Сейчас он содержит ∼1200 часов аудиодорожек и ∼600.000 высказываний. Тексты были взяты из книг, википедии, новостей и тд. Записи были сделаны на специальной платформе людьми разного пола, возраста, регионов проживания. Дополнительно какое-то количество аудиозаписей были провалидированы специально обученными людьми. Если данных будет не достаточно, то можно будет попробовать применить разные аугментации.
Так же нам может понадобится OpenSTT, Golos, Mozilla Common Voice с аудиозаписями на русском языке.

3 План

Я бы сделал так:

- Обучил обычную n-gram LM на казахских текстах
- Так как у нас относительно немного данных, то и модель нам нужна не слишком большая, но умная. На ум приходит QuartzNet. Как бейзлайн. Имплементацию сети можно найти, поэтому не придется писать код. Можно будет попробовать взять также Conformer или Jasper.
- В качестве словаря предлагаю попробовать 2 ранее описанных способа объединения алфавитов.
- Тренировать модели предлагаю сначала только на казахском датасете(моноязычная модель), а потом на обоих(казахском и русском) . Можно попробовать не тренировать модель с нуля, а инициализировать претрейн чекпоинтом другого языка - тогда она сразу будет иметь представление о звуках.
- Далее попробовал бы реализовать подход авторов последней статьи. Он выглядит интересным. Здесь уже придется писать код самим.
- Еще раз попробовать учить на BPE