Использование алгоритмов динамического программирования для анализа фонетической интерференции в китайско-русской речи

Максимова Яна Дмитриевна
Московский государственный университет
имени М.В. Ломоносова
s02210457@gse.cs.msu.ru

Аннотация

В данной работе исследуется фонетическая интерференция в китайско-русской речи с использованием методов динамической трансформации временной шкалы (dynamic time warping, dtw) и главных компонент сигнала. Целью работы является выявление правильного и неправильного произношения слов русского языка у носителей китайского языка для дальнейшего применения алгоритмов в области автоматического распознавания речи и обучения русскому языку как иностранному.

Keywords распознавание речи · обработка речевых сигналов · DTW-классфикация

1 Введение

Одной из актуальных задач современной фонетической науки является анализ фонетической интерференции в речи, особенно в контексте изучения иностранного языка. В случае китайского языка, который существенно отличается от русского по фонетическим особенностям, носители китайского языка сталкиваются с определенными трудностями при произнесении русских слов. Эти трудности проявляются в изменении или замене фонем, что влияет на правильность произношения.

В последние годы с развитием технологий машинного обучения и обработки речи возникает потребность в автоматизации процесса выявления таких ошибок. Это особенно важно для задач автоматического распознавания речи (ASR) и обучения иностранному языку, где необходимо точно определять, какие слова были произнесены правильно, а какие — с ошибками. Традиционные методы классификации, такие как динамическое программирование и алгоритм динамической трансформации временной шкалы (DTW), остаются актуальными, особенно в случаях, когда объем доступных данных ограничен.

В данной статье рассматривается применение методов динамического программирования для решения задачи классификации произнесенных слов и анализа фонетической интерференции в китайско-русской речи.

2 Постановка задачи

В рамках данной работы изучается применение алгоритма динамической трансформации временной шкалы (Dynamic Time Warping, DTW) для решения двух задач: первая задача заключается в классификации произнесений слов на правильные и неправильные, Вторая — в распознавании слова, произнесенного говорящим, при условии, что словарь возможных вариантов ограничен конечным множеством.

В рамках исследования было выделено 1500 слов, которые представляют наибольшую трудность для произношения у носителей китайского языка. В эксперименте принимали участие как мужчины, так и женщины, что добавило сложности к анализу из-за ограниченного количества участников.

Критерии оценки качества классификации правильности произнесенного слова будут включать метрики полноты, точности и F1-меры. Для второй задачи, связанной с распознаванием произнесенного слова, основными метриками качества выступают точность (accuracy) и средняя точность извлечения (mean average precision, mAP)

Решение должно соответствовать ряду требований:

- Алгоритм должен учитывать вариативность темпа и длительности произнесения слов разными участниками
- Эффективно справляться с вариативностью произношения, связанной с полом и индивидуальными особенностями произношения у экспериментуемых.
- Стабильность на ограниченном числе данных.
- Высокая точность классификации.

3 Решение

3.1 Описание алгоритма

Алгоритм динамической трансформации временной шкалы (DTW) представляет собой метод вычисления минимального расстояния между двумя временными последовательностями. Его основная особенность заключается в способности находить оптимальное соответствие между последовательностями, даже если они различаются по длине или имеют временные искажения. Это делает алгоритм особенно подходящим для задач сравнения акустических сигналов, где длина записей может значительно варьироваться.

Предположим, что Q и C — две временные последовательности длины n и m, соответственно, где [1]:

$$Q = q_1, q_2, \ldots, q_n \tag{1}$$

$$C = c_1, c_2, \ldots, c_m \tag{2}$$

Для совмещения этих двух последовательностей с использованием DTW строится матрица размерностью $n \times m$, где элемент на позиции (i, j) представляет расстояние $d(q_i, c_j)$ между точками q_i и c_j. Расстояние $d(q_i, c_j)$ определяется как евклидово и вычисляется по следующей формуле:

$$d(q_i, c_j) = (q_i - c_j)^2 \tag{3}$$

Элементы $d(q_i, c_j)$ называются локальными расстояниями. Каждый элемент (i, j) в матрице соответствует совмещению точек q_i и c_j. Процесс выравнивания осуществляется с помощью динамического программирования, что позволяет вычислить накопленное расстояние $g(i, j)$ по рекуррентной формуле:

$$g(i, j) = d(q_i, c_j) + \min\{g(i-1, j-1), g(i-1, j), g(i, j-1)\} \tag{4}$$

Если w обозначает траекторию выравнивания, то она должна удовлетворять следующим условиям:

- граничное условие: $w_1 = (1, 1)$ и $w_k = (m, n)$: требуемая траектория выравнивания такова, чтобы она начиналась первым и заканчивалась последним элементом главной диагонали,
- непрерывность: заданные $w_k = (a, b)$ и $w_{k-1} = (a', b')$, где $a - a' \leq 1$ и $b - b' \leq 1$. Что является ограничением разрешенной траектории выравнивания для соседних элементов,
- монотонность: заданные $w_k = (a, b)$ и $w_{k-1} = (a', b')$, где $a - a' \geq 0$ и $b - b' \geq 0$. Это требование монотонного увеличения во времени точек W. Например, нельзя отображать точку 6. точку 4, а затем точку 5

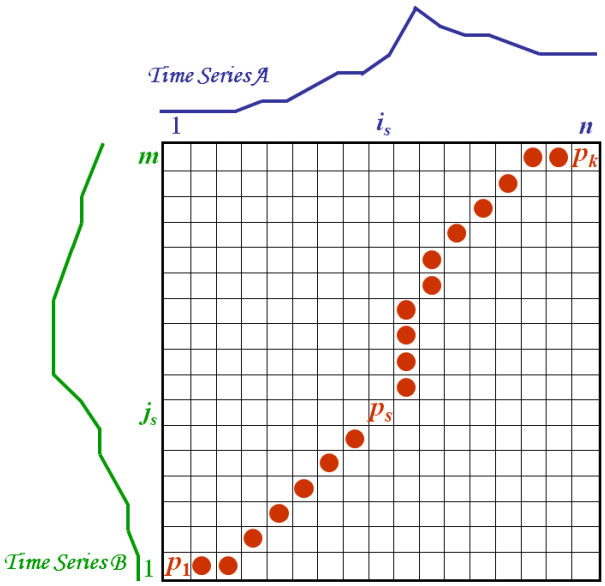

Рис. 1: Иллюстрация алгоритма выравнивания временных масштабов.

3.2 Извлечение признаков

Мел-спектральный анализ представляет собой преобразование звукового сигнала в частотную область с использованием мел-шкалы, которая учитывает особенности человеческого восприятия звука. В основе этого анализа лежит разложение сигнала на частотные компоненты с применением фильтров, равномерно распределенных в мел-шкале.

$$mel = 2695 \cdot \lg(1 + \frac{1}{700}) \tag{5}$$

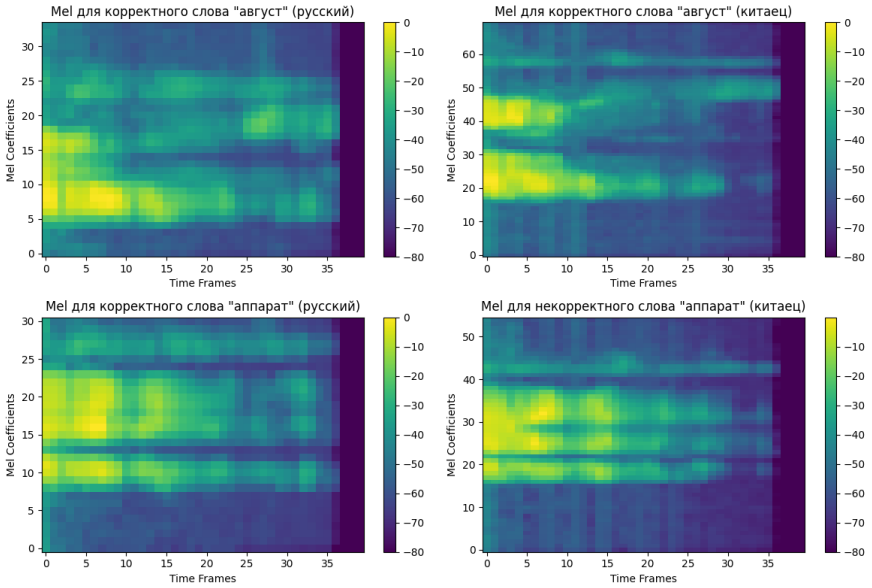

Рис. 2: Иллюстрация Мел-спектрограмм для корректного и некорректного произношения.

Мел-кепстральные коэффициенты (MFCC) представляют собой набор признаков, которые описывают форму спектральной огибающей звука, моделируя характеристики человеческого голоса. Эти

коэффициенты, отражая частотное распределение в пределах заданного временного окна, позволяют анализировать как частотные, так и временные аспекты звукового сигнала [2]. Для повышения устойчивости MFCC к различиям между голосами мужчин и женщин, перед их использованием выполняется нормализация.

$$c[n] = \sum_{k=1}^{K} \log(S[k]) \cdot \cos\left[n \cdot (k - 0.5) \cdot \frac{\pi}{K}\right], \tag{6}$$

где:

- n — номер коэффициента MFCC ($n = 1, 2, \ldots, N$),
- K — количество фильтров в мел-фильтробанке,
- $S[k]$ — энергия сигнала на выходе k-го фильтра в мел-фильтробанке,
- $\cos[\ldots]$ — косинусное преобразование, используемое для перехода от мел-спектрального представления к кепстральным коэффициентам.

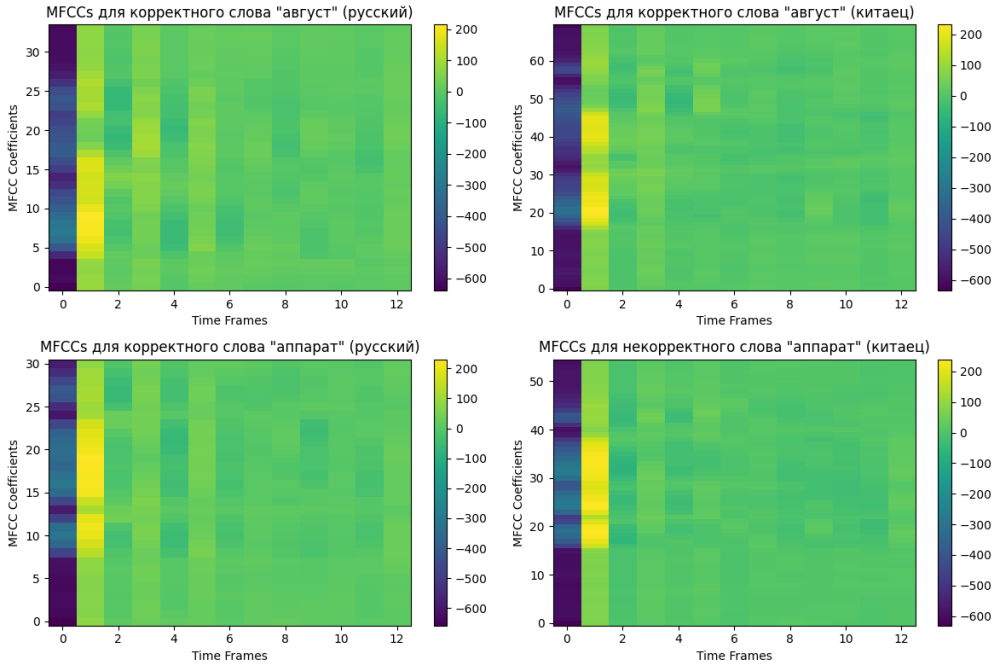

Рис. 3: Иллюстрация mfcc для корректного и некорректного произношения.

Как описано в статье [3], мел-спектральные признаки лучше подходят для задач общего распознавания речи, так как они фокусируются на частотных характеристиках. MFCC, обеспечивает более точное представление речи, что делает его предпочтительным в задачах классификации произношения и идентификации говорящего.

3.3 Альтернативные подходы решения задачи

Для сравнения с DTW можно рассмотреть альтернативные подходы, которые находят применение в задачах обработки речи. Одним из таких методов являются глубокие нейронные сети [4], включая рекуррентные архитектуры и трансформеры, которые способны обучаться на больших объемах данных [5, 4] и эффективно решать задачи распознавания и классификации речи. Также широко используются статистические модели, такие как скрытые марковские модели, которые традиционно применяются для анализа временных последовательностей [6].

Несмотря на достоинства и активное развитие данных подходов, DTW сохраняет свою актуальность для задач, где объем данных, как в нашем случае, сильно ограничен.

4 Вычислительные эксперименты

4.1 Цель эксперимента

Основной целью экспериментов является оценка эффективности алгоритма DTW в задачах классификации правильных и неправильных произнесений слов, а также в распознавании слов из конечного известного словаря.

4.2 Описание и постановка экспериментов

В ходе работы были проведены два типа экспериментов. Первый из них был направлен на классификацию произнесений слов. Для этого все записи были разделены на две группы: правильные и неправильные произнесения. Алгоритм DTW использовался для создания модели, способной различать эти категории. Метриками качества для оценки служили точность, полнота и F1-мера. Второй эксперимент касался задачи распознавания слов. Здесь каждая тестовая запись сравнивалась с эталонными (правильно произнесенными словами как носителей русского языка, так и китайского) реализациями из словаря, и слово классифицировалось на основе минимального значения DTW-расстояния. Для этой задачи оценивались точность и mAP.

В рамках дополнительных экспериментов была проведена оценка эффективности мел-спектральных и кепстральных признаков, а также исследована нормализация звуковых сигналов, что позволяет адаптировать модель для работы как с мужскими, так и с женскими голосами.

4.3 Описание данных

В эксперименте приняли участие 20 дикторов в возрасте от 20 до 25 лет: 10 носителей русского языка, среди которых было 4 женщины, и 10 носителей китайского языка, среди которых была 1 женщина. Каждый диктор произнес 1500 слов. При этом доля неправильно произнесенных слов среди носителей китайского языка составила от 70% до 80%.

Общая звуковая дорожка была разделена на отдельные файлы формата .wav с помощью сегментации, основанной на предварительно составленной разметке. Для проведения экспериментов использовались аудиофайлы, содержащие произнесения слов ("dictorX.wav где X – фамилия диктора) и соответствующие им файлы разметки ("dictorX.mrk"). Разметка включала текстовые метки слов, а также информацию о параметрах дикторов (пол, возраст, устройство записи) и качестве произнесения. Последний параметр обозначался как PronQuality=0 для неправильного произнесения или PronQuality=1 для правильного.

4.4 Описание хода экспериментов

На первом этапе проводилась предобработка данных. Длинные аудиофайлы сегментировались на отдельные слова на основе информации из файлов разметки. Для каждого сегмента вычислялись мел-спектральные и кепстральные признаки, которые служили основой для сравнения с использованием DTW. Все аудиоданные нормализовались, чтобы обеспечить сопоставимость характеристик.

В процессе обучения алгоритм DTW использовался для вычисления расстояний между различными произнесениями одного и того же слова. Для правильных и неправильных реализаций формировались два множества расстояний, которые затем использовались для определения порогов классификации. Во время тестирования каждая новая реализация сравнивалась с эталонными записями, а классификация выполнялась на основе минимального расстояния.

В качестве альтернативы данному методу можно рассмотреть использование модели wav2vec. Эта модель, основанная на глубоком обучении, применяет нейронные сети для извлечения признаков из аудиофайлов и используется для решения задач распознавания речи.

4.5 Результаты экспериментов

Ранее проведенные эксперименты показали следующие результаты. Распознавание слов, произнесенных носителями русского языка, оказалось на высоком уровне для обеих моделей. Однако модель wav2vec показала лучшие результаты, поскольку алгоритм, основанный на DTW, ограничен только нашими 1500 словами, и для масштабирования требуется переобучение на новом наборе данных. В то же время результаты распознавания русских слов, произнесенных носителями китайского языка, продемонстриро-

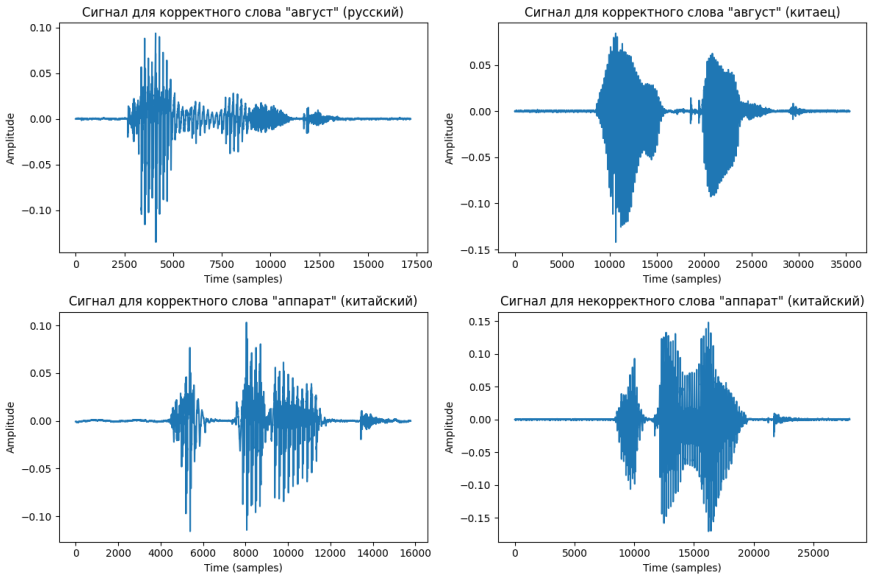

Рис. 4: Иллюстрация сигналов корректного и некорректного произношения.

вали явные преимущества нашего алгоритма. Он более чувствителен и не пытается подбирать аналоги слов, а ищет их среди ограниченного множества, что позволяет добиваться более точных результатов.

Таблица 1: Результат распознавания слов для носителей русского языка

	recall	*precision*	*F*1
wav2vec	0.915	0.864	0.88
dtw	0.753	0.826	0.78

Таблица 2: Результат распознавания слов для носителей китайского языка

	recall	*precision*	*F*1
wav2vec	0.231	0.422	0.29
dtw	0.673	0.520	0.54

5 Выводы

Алгоритм DTW зарекомендовал себя как надежный инструмент для анализа акустических данных, показывая высокую точность в задачах классификации и распознавания. Его устойчивость к временным искажениям делает его особенно подходящим для обработки вариативных речевых данных.

Список литературы

[1] И.А. Гуртуева and К.Ч. Бжихатлов. Аналитический обзор и классификация методов выделения признаков акустического сигнала в речевых системах. 2022.

[2] Hsiao-Wuen Hon, Сюэдун Huang, and Алехандро Acero. Spoken Language Processing, A Guide to Theory, Algorithm and System Development. 2001. URL `https://drive.google.com/file/d/11DMkh7N-6CuuzwOc8byBasBcxsYLkhWI/view`.

[3] Д.А. Кацай, Т.Д. Исупова, and Е.В. Харченко. Алгоритм анализа речевых образцов иностранных студентов, изучающих русский язык. 2018.

[4] Peter Bell, Joachim Fainberg, Ondrej Klejch, Jinyu Li, Steve Renals, and Pawel Swietojanski. Adaptation algorithms for neural network-based speech recognition: An overview. 2020.

[5] А.К. Алимурадов and П.П. Чураков. Обзор и классификация методов обработки речевых сигналов в системах распознавания речи. 2016.

[6] А.В. Малышев. Обзор технологий генерации и распознавания речи. 2024.