Machine Translated by Google

прикладные науки

MDPI

Статья

# Использование трансферного обучения для реализации дунган с низкими ресурсами Синтез языковой речи

Мэнгруй Лю 1, †, Руй Цзян 2, † и Хунву Ян 2,3,*

1 Колледж электронной и информационной инженерии, Университет Тунцзи, Шанхай 201804, Китай; liuxh709@163.com

2 Школа образовательных технологий Северо-Западного педагогического университета, Ланьчжоу 730070, Китай; jiangh940618@163.com

3 Ключевая лаборатория цифровизации образования провинции Ганьсу, Ланьчжоу 730070, Китай

* Переписка: yanghw@nwnu.edu.cn

† Эти авторы внесли одинаковый вклад в эту работу.

Аннотация: В данной статье представлен метод на основе трансферного обучения для улучшения качества синтезированной речи малоресурсного дунганского языка. Это улучшение достигается за счет тонкой настройки предварительно обученной акустической модели мандаринского языка на акустическую модель дунганского языка с использованием ограниченного корпуса дунганского языка в рамках Tacotron2+WaveRNN. Наш метод начинается с разработки анализатора дунганского текста на основе преобразователя, способного генерировать последовательности единиц со встроенной просодической информацией из дунганских предложений. Эти последовательности единиц, наряду с речевыми характеристиками, предоставляют пары <последовательность единиц с просодическими метками, Mel-спектрограммы> в качестве входных данных Tacotron2 для обучения акустической модели. Одновременно мы предварительно обучили акустическую модель мандаринского языка на основе Tacotron2, используя крупномасштабный корпус мандаринского языка. Затем модель дорабатывается с помощью небольшого дунганского речевого корпуса для получения дунганской акустической модели, которая автономно изучает выравнивание и сопоставление единиц со спектрограммами. Полученные спектрограммы преобразуются в сигналы с помощью вокодера WaveRNN, что облегчает синтез высококачественной мандаринской или дунганской речи. Как субъективные, так и объективные эксперименты показывают, что предлагаемый синтез дунганской речи на основе трансферного обучения дает более высокие результаты по сравнению с моделями, обученными только с использованием дунганского корпуса и других методов. Следовательно, наш метод предлагает стратегию достижения синтеза речи для языков с низкими ресурсами путем добавления просодической информации и использования аналогичного языкового корпуса с высокими ресурсами посредством трансферного обучения.

Ключевые слова: синтез речи дунганского языка; анализ текста; трансферное обучение; малоресурсный язык; такотрон2

check for updates

Цитирование: Лю, М.; Цзян, Р.; Ян, Х. Использование трансферного обучения для реализации синтеза речи на дунганском языке с низкими ресурсами. Прил. наук. 2024, 14, 6336. https://doi.org/10.3390/app14146336 .

Академические редакторы: Глория Корпус Пастор и Таринду Ранасингхе

Поступила: 17 июня 2024 г.

Пересмотрено: 17 июля 2024 г.

Принято: 18 июля 2024 г.

Опубликовано: 20 июля 2024 г.

Копирайт: © 2024 авторов. Лицензиат MDPI, Базель, Швейцария.

Эта статья находится в открытом доступе. распространяется на условиях и условия Creative Commons

Лицензия с указанием авторства (CC BY) ( https://creativecommons.org/licenses/by/4.0/).

## 1. Введение

Синтез речи (преобразование текста в речь (TTS)) широко используется в умных домах, навигационных системах и приложениях для аудиокниг. В мире существует около 6000 языков , большинство из которых считаются малоресурсными. Хотя значительный прогресс был достигнут в синтезе речи для основных языков, таких как мандаринский, английский и французский, качество речи TTS для языков с ограниченными ресурсами, таких как тибетский и дунганский, остается неоптимальным. В последние годы наблюдается всплеск исследований, посвященных синтезу речи малоресурсных языков, о чем свидетельствуют многочисленные исследования [1–6]. Однако исследования по синтезу речи дунганского языка еще предстоит завершить. Дунганский язык, являющийся вариантом шаньси -ганьсуйских диалектов в составе китайского диалекта, на котором говорят в Центральной Азии, классифицируется как малоресурсный язык из-за ограниченности его использования, сокращения числа носителей и скудности лингвистического материала [7, 8]. Учитывая, что русский язык стал официальным языком Центральной Азии, создание комплексного речевого корпуса с лингвистическими знаниями для качественного синтеза дунганской речи представляет соб

Machine Translated by Google



Синтез дунганской речи на основе DNN [9,10] качество синтезированной речи было невысоким из-за ограниченности обучающего корпуса.

Технологии синтеза речи включают конкатенативный синтез речи на основе выбора единиц [11], статистический параметрический синтез речи на основе скрытой модели Маркова (HMM) (SPSS) [12] и синтез речи на основе глубокого обучения [13,14]. В то время как глубокое обучение значительно продвинуло технологию синтеза речи, такие методы, как длинная кратковременная память (LSTM) и двунаправленный LSTM [15,16], позволили устранить ограничения временной информации. Более того, модели сквозного синтеза речи [17], такие как Tacotron [18] и Tacotron2 [19], продемонстрировали способность напрямую отображать текст в речь. При обучении на крупномасштабных парах преобразования текста в речь эти модели производят синтезированную речь с использованием высококачественных вокодеров, таких как алгоритм Гриффина-Лима [20], WaveNet [21] и WaveRNN [22]. Однако такие системы требуют значительного обучения. Для малоресурсных языков отсутствие обучающего корпуса затрудняет изучение просодической структуры предложений сквозными моделями, что приводит к отсутствию просодических изменений в синтезируемой речи, что влияет на ее естественность, создавая проблемы для синтеза речи. языков с низкими ресурсами.

Межъязыковое трансферное обучение [23–25] использовалось для смягчения проблемы недостаточности обучающих корпусов для синтеза речи на языках с ограниченными ресурсами. Этот метод предполагает обучение языковой модели с использованием комбинации большого корпуса языка с высокими ресурсами и меньшего корпуса языка с низкими ресурсами с последующей адаптацией этой модели к языку с низкими ресурсами. Трансферное обучение в синтезе речи оказалось эффективной стратегией создания речи на языках с низким уровнем ресурсов за счет использования возможностей акустической модели языка с высокими ресурсами [26,27].

В нашем предыдущем исследовании синтеза тибетской речи [28–32] мы определили, что интеграция просодической информации с помощью методов трансферного обучения повышает качество синтезированной речи для языков с низкими ресурсами, таких как тибетский. Основываясь на этом понимании, настоящее исследование реализует подход «последовательность-последовательность» (seq2seq) для синтеза речи на дунганском языке, используя трансферное обучение и просодическую информацию в рамках Tacotron2 + WaveRNN. Этот метод предполагает использование анализатора текста дунганского языка для извлечения просодических меток из предложений дунганского языка для интеграции модели, использование акустической модели мандаринского языка на основе Tacotron2 и тонкую настройку акустической модели дунганского языка с ограниченным корпусом дунганской речи. Основные вклады представлены ниже:

- Интерфейс: мы реализовали полноценный анализатор текста для дунганского языка, включающий модули для нормализации текста, сегментации слов, просодического предсказания границ и генерации единиц измерения на основе технологии преобразования. Этот анализатор может формировать инициалы и финалы как единицы синтеза речи с просодическими метками из дунганских предложений.
- Серверная часть: мы добились синтеза речи на дунганском языке seq2seq, адаптировав предварительно обученную акустическую модель китайского языка в рамках Tacotron2+WaveRNN. Это было достигнуто за счет замены внимания Tacotron2, чувствительного к местоположению, на внимание вперед, что повысило скорость и стабильность конвергенции.

Остальная часть статьи организована следующим образом. Впервые мы представляем нашу структуру синтеза дунганской речи на основе трансферного обучения под управлением Tacotron2+WaveRNN в разделе 2. Экспериментальная установка и результаты представлены в разделе 3, а результаты обсуждаются в разделе 4. Наконец, краткий вывод и план будущей работы. представлены в раз

## 2. Модели и методы

Предлагаемая структура синтеза дунганской речи с низкими ресурсами на основе трансферного обучения, показанная на рисунке 1, включает модуль извлечения признаков, предварительно обученную акустическую модель мандаринского языка, обучающий модуль дунганской акустической модели на основе трансферного обучения и синтезатор речи на основе вокодера WaveRNN.

Machine Translated by Google



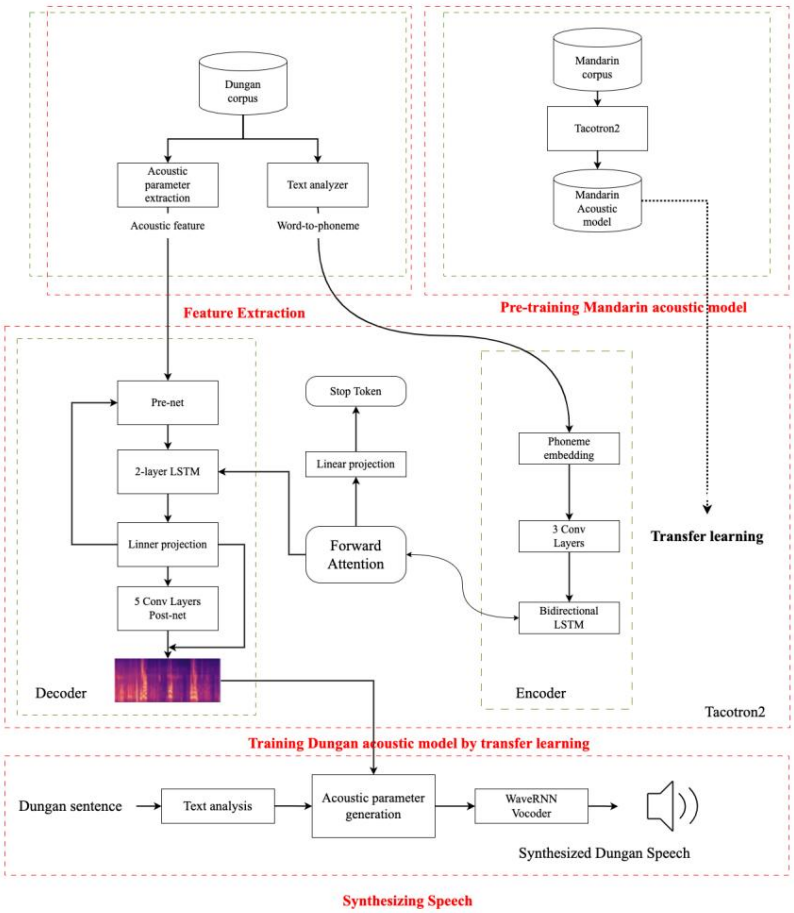


Рисунок 1. Схема синтеза дунганской речи на основе Tacotron2+WaveRNN.

Модуль извлечения признаков извлекает акустические характеристики, такие как мел-спектрограмма, из речевых сигналов и последовательность блоков синтеза речи из предложений. Мы разработали полноценный анализатор текста на дунганском языке для извлечения единиц синтеза речи с просодическими характеристиками и сопоставления дунганских предложений с последовательностями единиц. Учитывая, что и мандаринский, и дунганский языки используют инициалы и финалы в качестве основных единиц синтеза речи, результирующая последовательность единиц включает в себя эти элементы и соответствующую просодическую информацию, включая слоговые тона и просодические граничные метки на уровне предложения.

Поскольку Tacotron2 является одной из самых популярных платформ синтеза речи между кодерами и декодерами, а вокодер WaveRNN может генерировать естественную речь, мы используем Tacotron2 для обучения акустических моделей и WAVRNN для преобразования спектрограммы в форму волны как для дунганского языка, так и для мандаринского языка. Акустическая модель мандаринского языка предварительно обучается с помощью крупномасштабного корпуса мандаринского языка, а модель дунганского языка переносится из акустической модели мандаринского языка с мелкомасштабным корпусом дунганского языка.

На этапе синтеза речи вокодер WaveRNN генерирует дунганскую или мандаринскую речь на основе входных предложений дунганского или китайского языка. Анализатор текста сначала генерирует контекстно-зависимые метки из входного предложения. Затем последовательности единиц синтеза речи (инициалы и финалы с их просодической информацией) вводятся в акустическую модель мандаринского или дунганского языка для создания спектрограммы Мела. Вокодер WaveRNN наконец используется для генерации речевых сигналов из мел-спектрограммы. Для анализа китайского текста мы используем отечественный анализатор китайского текста.

### 2.1. Анализатор текста дунганского языка

В отличие от распространенных методов синтеза речи seq2seq, разработанных для основных языков, которые используют исключительно пару <последовательность фонем, речь> для обучения акустических моделей, наш подход использует последовательность единиц, включающую просодические метки, такие как

Machine Translated by Google



тон каждого слога и просодическая граница предложения, служащая «последовательностью фонем». Следовательно, становится необходимым разработать комплексный анализатор текста, способный извлекать последовательности единиц предложения и их просодические метки. С этой целью, используя наш собственный анализатор китайского текста, мы разработали анализатор текста на дунганском языке, как показано на рисунке 2. Процесс начинается с нормализации и сегментации входного дунганского предложения для определения границы слова. За этим следует анализ просодических границ, чтобы определить границу как просодического слова, так и просодической фразы. На заключительном этапе инициалы и финалы дунганских символов получаются посредством процесса преобразования символов в единицы измерения на основе преобразователя.

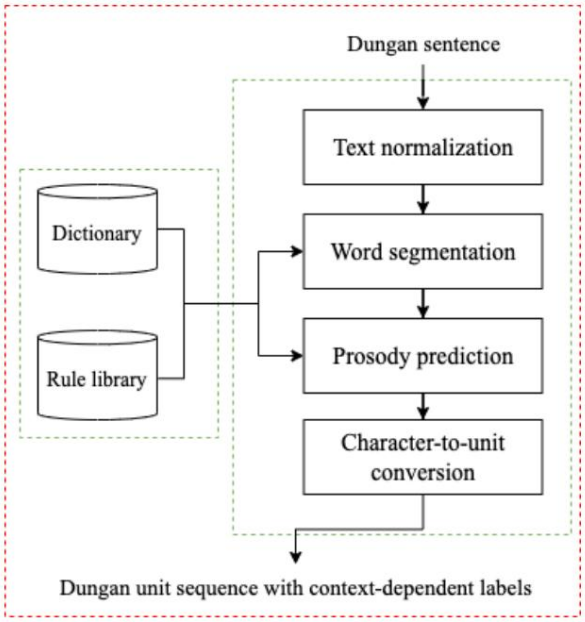


Рисунок 2. Процедура анализа дунганского текста.

### 2.1.1. Блок синтеза речи дунганского языка

Несмотря на использование другой системы письма, дунган представляет собой диалектное произношение мандаринского языка за пределами Китая. Дунганский язык написан кириллицей, напоминающей славянские языки, такие как русский, поэтому дунганский язык представляет собой фонетические символы с последовательным написанием, имеющие структуру, аналогичную китайской [33–35]. Порядок написания дунганских символов состоит из инициалов, финалов и тона, как показано на рисунке

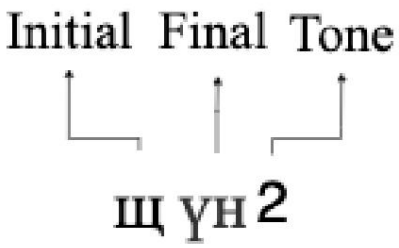


Рисунок 3. Структура дунганского характера.

В данной статье в качестве единицы синтеза речи используются инициалы и финалы. Дунганский иероглиф состоит из 25 инициалов (включая нулевой инициал) и 32 финалов, как показано в таблице 1. Как и мандаринский язык, тона дунганского языка имеют решающее значение для различения семантики и эмоций [36]. В дунгане четыре тона, за исключением светлого тона, а именно ровный тон (21), восходящий тон (24), ниспадающий тон (53) и нисходящий тон (44), каждый из которых обозначается цифрами от 1 до 4 соответственно.

Machine Translated by Google



Таблица 1. Инициалы и финалы дунганского языка.

| | |
|---|---|
| инициалы | /б/, /п/, /м/, /ф/, /в/, /з/, /ц/, /с/, /д/, /т/, /н/, /л/ /ж/ , /ch/, /sh/, /r/, /j/, / q/, /x/, /g/, /k/, /ng/, /h/, /φ/ /ii/, /iii /, /i/, /u/, /y/, /a/, /ia/, /ua/, /e/, /ue/, /ye/, /iE/ / |
| финал | ap/, /ai, /uai /, /ei/, /ui/, /ao/, /iao/, /ou/, /iou/, /an/, /ian/ /uan/, /yan/, /aN/, /iaN/, / yaH/, /yH/, /иH/, /yH/ |

2.1.2. Нормализация текста

Любое входное предложение может содержать числовые формы времени, даты, сокращения и специальные собственные существительные. Прежде чем преобразовать предложение в последовательность фонетических символов, необходимо использовать нормализацию текста, чтобы преобразовать нестандартный текст в единый фонетический символ. Поэтому мы реализовали нормализацию текста на основе правил для идентификации недунганских символов. Мы разработали набор правил нормализации дунганского текста на основе правил нормализации китайского текста [37] и применили метод добавления-восстановления для нормализации дунганских символов в соответствии с [38].

2.1.3. Сегментация слов

Границы слов играют важную роль в прогнозировании просодических границ. Таким образом, важно определить границы слов предложения после нормализации. Дунганские предложения демонстрируют четкие различия между словами и слогами, что делает сегментацию относительно простой. Мы использовали алгоритм сегментации слов на основе максимального соответствия для извлечения дунганских слов из входного предложения. Чтобы облегчить этот процесс, мы составили словарь дунганских слов , содержащий 49 293 слова. Самое длинное слово в этом словаре состоит из восьми символов, а самое короткое — из одного символа. Словарь в основном охватывает основные дунганские термины, на которые есть ссылки в таких источниках, как «Общий словарь дунганского языка» [39], «Обзор тунганского языка в Центральной Азии» [40], «Обзор дунганского языка» [41] и дополнительные дунганские термины, доступные для поиска, доступны в Интернете.

2.1.4. Прогнозирование просодической

границы Наш подход использует инициалы и финалы вместе с их просодическими метками в качестве входной последовательности единиц акустической модели. Таким образом, извлечение просодической структуры из дунганских предложений имеет решающее значение для синтеза качественной речи. Как и мандаринский язык, просодическая иерархия дунганского языка может быть разделена на просодические слова, просодические фразы, интонационные фразы и паузы в предложениях. Границу интонационных фраз легко определить с помощью дунганских знаков препинания. В этом исследовании мы использовали BiLSTM с методом на основе условного случайного поля (BiLSTM_CRF), как показано на рисунке 4, для прогнозирования границ просодических слов и фраз [42].

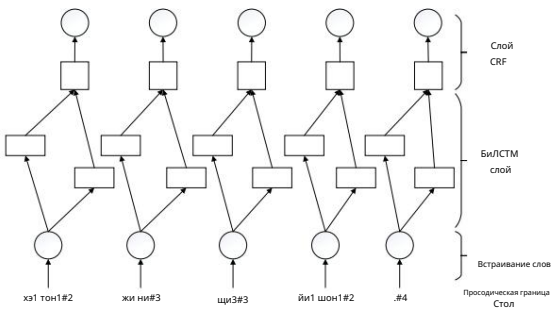


Рисунок 4. Схема прогнозирования дунганской просодической границы на основе BLSTM_CRF. Входные данные — дунганское предложение с просодической информацией.

Мы использовали четыре различных набора маркировок просодических позиций слов (№1, №2, №3, №4), чтобы классифицировать дунганские слова на просодические фразы. В частности, № 1 использовался для обозначения просодических слов, № 2 обозначал просодические фразы, № 3 обозначал окончание дунганского языка.

Machine Translated by Google



слово, а № 4 обозначал паузу в предложении. Процесс маркировки включал фразовую и просодическую информацию, полученную из дунганского текста, который был помечен вручную. На этом этапе лингвисты время от времени пересматривали и исправляли отдельные предложения. Мы достигли высокого уровня согласованности с экспертами-лингвистами посредством итеративных исправлений.

Несмотря на способность BiLSTM изучать контекстно-зависимую информацию, его независимые решения по классификации ограничены сильными зависимостями по выходной метке.

Чтобы решить эту проблему, мы используем уровень CRF, который учитывает соседние теги, как показано на рисунке 4. Для нормализованного входного предложения X = {x1, x2, · · · , xn} , содержащего n слов и последовательность тегов предложения y = ( y1, y2, . , yn), каждое слово представляется в виде d-мерного вектора word2vec. Мы определяем его оценку прогнозирования s(X, y) следующим образом:

$$s(X, y) = \sum_{i=1}^{n} P_{i,y_i} + \sum_{i=0}^{n} A_{y_i, y_{i+1}} \quad (1)$$

где P — матрица оценок, выдаваемых сетью BLSTM. Pi,yi соответствует баллу тега yi i-го слова в предложении. A — матрица оценок перехода уровня CRF, а Ayi ,yi+1 соответствует оценке от тега yi к тегу yi+1 .

В процессе обучения мы максимизируем следующие функции логарифмического правдоподобия:

$$\log(p(y \mid X)) = s(X, y) - \log \sum_{y \in Y_X} e^{c(X,y)} \quad (2)$$

где YX представляет все возможные последовательности тегов для входного текста X.

При декодировании оптимальная последовательность y задается следующим образом:

$$y^{*} = \operatorname{argmax}_{y \in Y_X} s(X, y) \quad (3)$$

2.1.5. Преобразование символов в единицы измерения на основе преобразователя

Мандаринский и дунганский языки используют одну и ту же систему пиньинь для обозначения произношения. Следовательно, преобразование символов в единицы в дунгане аналогично преобразованию в китайском языке. В этом исследовании представлен подход на основе преобразователя [43] для получения дунганской единицы, как показано на рисунке 5, для повышения точности преобразования дунганских символов в единицы. Кодер и декодер формируются путем объединения одних и тех же основных уровней с N = 6. Каждый базовый уровень состоит из двух подуровней. Первый подуровень — это уровень многоголового внимания. В декодере имеется слой скрытого многоголового внимания (маскированного многоголового внимания).

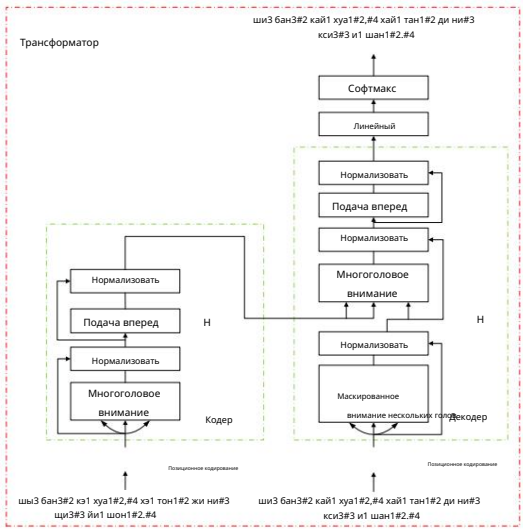


Рисунок 5. Схема преобразования дунганских символов в единицы на основе Transformer. Входными данными являются дунганское предложение с просодической информацией (слева) и соответствующей ему последовательностью пиньинь (справа). Результатом является последовательность пиньинь с просодической информацией.

Machine Translated by Google



2.2. Дунганская акустическая модель на основе трансферного обучения.

Мы реализуем дунганскую акустическую модель путем точной настройки предварительно обученного мандаринского языка. акустическая модель, как показано на рисунке 6.

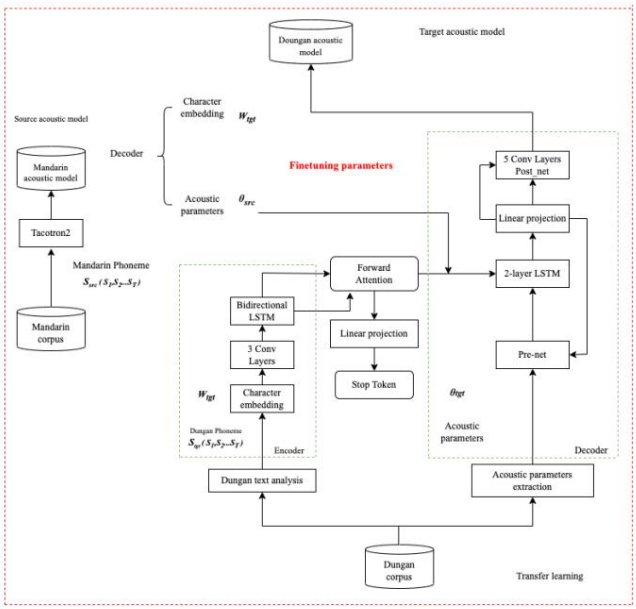


Рисунок 6. Процедура обучения акустической модели дунганского языка методом трансферного обучения.

2.3. Предварительно обученная акустическая модель китайского языка на основе Tacotron2

Акустическая модель китайского языка изначально обучается с использованием крупномасштабного корпуса китайского языка. Наш собственный анализатор китайского текста извлекает инициалы и финалы этих предложений, а также соответствующие просодические метки. Извлеченные акустические характеристики включают в себя мел-спектрограмму крупномасштабного корпуса мандаринского языка в рамках Tacotron2.

Учитывая схожее произношение дунганского языка и мандаринского языка, мы используем метод обучения картографирования-переноса [44] для получения акустической модели дунганского (целевого языка) путем передачи знаний с мандаринского языка (исходного языка), которую можно сформулировать следующим образом:

$$f\theta,W : XL \quad Y \tag{4}$$

где θ — параметры акустической модели, W обозначает встраивания обучаемых символов, а Y представляет собой пространство мандаринского языка. XL — текстовое пространство дунганского языка.

$$XL = \{ст\} \quad \begin{matrix} T \\ T=1 \end{matrix} \quad | \quad tst \quad L, T \quad N \tag{5}$$

где L — набор единиц дунганского языка, St — t-я единица последовательности единиц дунганского языка, а T — длина последовательности единиц.

В кодировщик мы вводим последовательность дунганских единиц, представленную встраиванием символов. Он проходит через стек из трех сверточных слоев, за которым следует пакетная нормализация и активации ReLU. Впоследствии выходные данные окончательного сверточного слоя передаются в двунаправленный слой LSTM для генерации функций дунганского блока.

Переносное обучение на основе картирования включает в себя отображение экземпляров из θsrc и θtgt в новое пространство акустических параметров. В этом процессе мы можем напрямую использовать Wsrc и θsrc, декодированные декодером из акустической модели китайского языка. θsrc и θtgt могут принимать вложения в качестве входных данных и генерировать речь. Однако, поскольку ssrc и stgt происходят из разных наборов символов, т. е. Lsrc = Ltgt, одну и ту же концепцию нельзя напрямую применить к Wsrc и Wtgt. Чтобы решить эту проблему, в Wtgt встроены единицы дунггана, облегчающие переобучение в процессе передачи.

Мы применяем механизм прямого внимания, который использует совокупные веса внимания. для вычисления вектора контекста.

Machine Translated by Google



Декодер представляет собой авторегрессионную рекуррентную нейронную сеть, которая прогнозирует θtgt из кодер вводит дунганскую единицу последовательности по одному кадру за раз. Мы можем использовать изученное θsrc из акустической модели китайского языка для инициализации θtgt в новом пространстве акустических параметров. Выходные данные начального временного шага сначала обрабатываются через Pre-net, состоящую из два полностью связанных слоя. Этот результат сочетается с контекстом прямого внимания. вектор и прошел через пару слоев LSTM. Комбинация выходов LSTM и векторы контекста внимания подвергаются трем различным линейным преобразованиям, чтобы предсказать целевой кадр спектрограммы, стоп-токен и предполагаемый остаток. Впоследствии прогнозируемое акустические характеристики подвергаются пяти сверточным слоям, генерируя остаток для улучшения реконструкция акустической модели дунганов.

## 3. Результаты

### 3.1. Оценка преобразования дунганских символов в единицы на базе трансформатора

Анализ текста во фронтенде влияет на качество синтеза речи сзади. конец, поэтому мы оценили дунганский анализатор текста, в котором преобразование символов в единицы Модуль является наиболее критичным фактором, влияющим на качество синтезируемой речи. Оценивать жизнеспособность дунганского модуля преобразования символов в единицы на основе трансформатора, мы использовали набор данных, содержащий 10 783 предложения на дунганском языке, расшифрованные с помощью Мандаринский пиньинь. Дунганский язык набора данных и представления мандаринского пиньинь являются изоморфными, заключающими в себе текстовые атрибуты, такие как тон и просодические границы, присущие на дунганский язык. В нашем исследовании мы выделили 10% из 10 783 предложений. в качестве тестового набора, еще 10% — в качестве набора для проверки, а оставшиеся 80% были обозначен как обучающий набор. Гиперпараметры, связанные с Трансформатором: подробно описано в Таблице 2. В качестве показателей оценки мы использовали показатели точности, полноты и F1. показано в Таблице 3. Результаты процесса оценки подтвердили, что предлагаемый Дунганский модуль преобразования символов в единицы подходит для последующей оценки синтеза речи.

Таблица 2. Гиперпараметры модели преобразования символов в единицы на основе преобразователя.

| Параметр | Ценить |
|---|---|
| Слои внимания Nx | 6 |
| Руководители | 8 |
| Размер партии | 32 |
| Скрытый | 513 |
| Выбывать | 0,1 |
| Скорость обучения | 0,0001 |

Таблица 3. Результаты преобразования дунганских символов в единицы на основе Трансформатора.

| Точность | Отзывать | Ф1 |
|---|---|---|
| 90.12 | 89,91 | 90.01 |

### 3.2. Оценка дунганских акустических моделей на основе трансферного обучения

#### 3.2.1. Корпус

В эксперименте мы использовали записи девяти женщин и тридцати одного мужчины. из 30-часовой базы данных китайского языка Цинхуа [45] (всего 13 389 предложений) как корпус мандаринского языка. Для дунганского корпуса мы выбрали пять записей мужчин (923 на каждого). человек, всего 4615 предложений и 6 ч). Дунганский корпус охватывает все начальные и окончательные произношения дунганского языка. Средняя длина предложения составляет 18 слогов, средней продолжительностью 10 с. Все записи были конвертированы в моноканальный 16 кГц. частота дискретизации с точностью квантования 16 бит.

Machine Translated by Google



3.2.2. Экспериментальная установка

Три типа фреймворков TTS, включая Tacotron+Griffin-Lim, Tacotron2+WaveNet, и Tacotron2+WaveRNN сравнивались в экспериментах. Некоторые гиперпараметры схемы представлены в Таблице 4.

Таблица 4. Гиперпараметры моделей Tacotron и Tacotron2.

| Модель | | Такотрон | Такотрон2 | Такотрон2 с передовым вниманием |
|---|---|---|---|---|
| Вокодер | | Гриффин-Лим | ВейвНет | ВолнаRNN |
| Кодер | Встраивание | Фомема (256) | Фомема (512) | Фомема (512) |
| | Предварительная сеть | ФФН (256, 128) | - | ФФН Фомема (512, 256) |
| | Ядро энкодера | ЦБХГ (256) | Си-Эн-Эн (512) Би-LSTM (512) | Си-Эн-Эн (256) Би-LSTM (256, 512) |
| Декодер | Пост-нет | ЦБХГ (256) | Си-Эн-Эн (512) | Си-Эн-Эн (512) |
| | Декодер РНН | ГРУ (256, 256) | - | ЛСТМ (512, 256) |
| | Внимание | Добавка (256) | чувствительный к местоположению (128) | Нападающий (256) |
| | Внимание РНН | ГРУ (256) | ЛСТМ (1024, 1024) | ЛСТМ (256) |
| | Предварительная сеть | ФФН (256, 128) | ФФН (256, 256) | ФФН (256, 128) |
| Параметр | | 7,6 × 106 | 28,9 × 106 | 23,7 × 106 |

Все три фреймворка включают в себя модуль внешнего анализатора текста, акустическую модель обучающий модуль и вокодер. Модуль анализатора текста преобразует дунганский или китайский языки. предложения в последовательность единиц, представленную пиньинь, включая инициалы, финалы и их тона и просодические метки границ. В модуле обучения акустической модели мы получаем лог Спектрограмма магнитуды речевого сигнала с использованием окна Ханна с периодом 80 мс длина кадра, сдвиг кадра 12,5 мс и преобразование Фурье по 2048 точкам.

Для платформы Tacotron+Griffin-Lim акустические модели обучаются с использованием выходных данных. коэффициент уменьшения слоя r = 3 и оптимизатор Адама с затухающей скоростью обучения. Скорость обучения начинается с 0,001 и впоследствии снижается до 0,0005, 0,0003 и 0,0001. через 5, 20 и 50 эпох соответственно. Для расчета используется прямая функция потерь. декодер seq2seq (Mel-спектрограмма) и сеть постобработки (линейная спектрограмма). Размер обучающего пакета установлен на 32, при этом все последовательности дополняются до максимальной длины на восстановление дополненных нулями кадров. В качестве вокодера используется алгоритм Гриффина-Лима. для преобразования спектра Mel в речь.

Для платформы Tacotron2+WaveNet мы обучаем акустические модели, используя стандартная процедура обучения максимального правдоподобия, которая включает в себя подачу правильных выходных данных вместо прогнозируемого вывода на стороне декодера. Это было завершено с размером партии из 32. Использовался оптимизатор Адама со следующими параметрами: β = 0,9, β = 0,999, = 10 6 . Скорость обучения была инициализирована на уровне 10-3 , а затем экспоненциально падала до 10-5. после 50 000. Дополнительно мы применили регуляризацию L2 с весом 10 6 . Для Мела Для преобразования спектра в речь в качестве вокодера использовалась WaveNet.

В нашей системе трансферного обучения на основе Tacotron2+WaveRNN мы изначально используем крупномасштабный корпус китайского языка для предварительной подготовки акустической модели китайского языка для последующей модели передача. Эта предварительно обученная модель затем используется для обучения дунганской акустической модели посредством передачи обучение по мандаринско-дунганскому корпусу. Для вокодирования мы используем WaveRNN для Преобразование мел-спектра в речь. Учитывая, что настройки параметров существенно влияют точность и надежность модели, мы оптимизировали эти параметры посредством итеративного обучения и обновления.

Каждая структура TTS реализует одноязычный синтез речи для мандаринского или дунганского языка и двуязычный синтез на основе трансферного обучения. Мы обучили несколько моделей в трех

Machine Translated by Google



Системы TTS для оценки качества и четкости синтезированной речи. В нашем эксперименте 10% высказываний были случайным образом распределены в тестовый набор, еще 10% были обозначены для развивающего набора, а остальные высказывания составляли обучающий набор.

Дунганская одноязычная модель, зависящая от говорящего

Мы обучили акустическую модель дунганской монолингвальной динамики (DSD), используя записи пяти говорящих мужчин, каждый из которых составил 923 предложения, всего 4615. предложения и охватывают 6 часов. Затем мы сравнили качество и четкость синтезированных речь в трех средах: DSD-Tacotron+Griffin-Lim, DSD Tacotron2+WaveNet, и DSD-Tacotron2+WaveRNN.

Мандаринская одноязычная модель, зависящая от говорящего

Мы использовали записи девяти женщин и тридцати одного мужчины (Цинхуа). 30-часовая база данных китайского языка, состоящая из 13 389 предложений) для обучения моноязычному китайскому языку. Акустическая модель, зависящая от динамика (MSD). Сравнили качество синтезированной речи и ясность в трех структурах: MSD-Tacotron+Griffin-Lim, MSD-Tacotron2+WaveNet, и MSD-Tacotron2+WaveRNN.

Двуязычная модель мандаринского и дунганского языков, зависящая от говорящего

Мы использовали записи пяти дунганцев-мужчин (923 предложения на человека, суммирование до 4615 предложений, что эквивалентно 6 часам) в качестве обучающих данных для перевода мандаринской акустической модели в дунганскую акустическую модель для реализации дунганской зависимости от говорящего. (MDSD) акустическая модель и акустическая модель, зависящая от говорящего на китайском языке (MDSM). Мы затем сравнили качество и четкость синтезированной речи в шести системах.

- MDSD-Такотрон+Гриффин-Лим
- МДСМ-Такотрон+Гриффин-Лим
- MDSD-Tacotron2+WaveNet
- МДСМ-Такотрон2+WaveNet
- MDSD-Такотрон2+WaveRNN
- МДСМ-Такотрон2+WaveRNN

3.2.3. Объективные оценки

Мы использовали Мел-кепстральное искажение (MCD) [46], Искажение периодичности полосы А. (BAP) [47], среднеквадратическая ошибка (RMSE) [48] и вокализованная/невокализованная ошибка (V/UV) [47] объективно оценить различные модели. Результаты для акустики DSD и MSD модели представлены в табл. 5 и табл. 6 соответственно. Аналогичным образом, MDSM и MDSD результаты показаны в Таблице 7 и Таблице 8 соответственно.

Таблица 5. Объективные результаты акустической модели DSD для Дунгани.

| Модель | Такотрон+Гриффин-Лим | Tacotron2+WaveNet | Tacotron2+WaveRNN |
|---|---|---|---|
| МКД (дБ) | 9,675 | 9.572 | 9.502 |
| БАД (дБ) | 0,189 | 0,187 | 0,170 |
| F0 RMSE (Гц) | 32,785 | 32,692 | 32.087 |
| В/УФ (%) | 9,867 | 9,721 | 9,875 |

Таблица 6. Объективные результаты акустической модели MSD для Мандарина.

| Модель | Такотрон+Гриффин-Лим | Tacotron2+WaveNet | Tacotron2+WaveRNN |
|---|---|---|---|
| МКД (дБ) | 5,460 | 5,291 | 5.036 |
| БАД (дБ) | 0,174 | 0,171 | 0,169 |
| F0 RMSE (Гц) | 14,629 | 13,986 | 13,647 |
| В/УФ (%) | 5,619 | 5,793 | 5,762 |

Machine Translated by Google



Таблица 7. Объективные результаты акустической модели MDSD для Дунгани.

| Модель | Такотрон+Гриффин-Лим | Tacotron2+WaveNet | Tacotron2+WaveRNN |
|---|---|---|---|
| МКД (дБ) | 7,523 | 7,419 | 7.395 |
| БАД (дБ) | 0,178 | 0,175 | 0,174 |
| F0 RMSE (Гц) | 26,891 | 26,753 | 26,617 |
| В/УФ (%) | 7,774 | 7,693 | 7.607 |

Таблица 8. Объективные результаты акустической модели MDSM для китайского языка.

| Модель | Такотрон+Гриффин-Лим | Tacotron2+WaveNet | Tacotron2+WaveRNN |
|---|---|---|---|
| МКД(дБ) | 5.339 | 5.241 | 5.108 |
| БАД (дБ) | 0,174 | 0,173 | 0,171 |
| F0 RMSE (Гц) | 13,775 | 13.326 | 13.092 |
| В/УФ (%) | 5.542 | 5.472 | 5.481 |

В условиях малоресурсного синтеза дунганской речи качество координации внимания между кодером и декодером существенно влияет на качество синтезируемой речи. речь. Несовпадения в первую очередь проявляются в читаемости, пропусках и повторении. Следовательно , мы используем коэффициент диагональной фокусировки (DFR) и уровень разборчивости на уровне слов. (IR) [49] для оценки читаемости на языках с ограниченными ресурсами, как показано в Таблице 9. DFR представляет собой карту внимания между кодером и декодером, служащую архитектурным метрика. IR измеряет процент тестовых слов, произнесенных правильно и четко. люди — стандартная метрика для оценки качества генерации речи с низкими ресурсами.

Таблица 9. Читабельность синтезированной дунганской речи.

| Модель | ИР (%) | ДФР (%) |
|---|---|---|
| DSD-Такотрон+Гриффин-Лим | 82,93 | 79,64 |
| DSD-Такотрон2+WaveNet | 86,67 | 82,43 |
| DSD-Такотрон2+WaveRNN | 89,41 | 84,39 |
| MDSD-Такотрон+Гриффин-Лим | 95.03 | 91,14 |
| MDSD-Такотрон2+WaveNet | 96,69 | 94,43 |
| MDSD-Такотрон2+WaveRNN | 98,47 | 97,39 |

3.2.4. Субъективная оценка

Для субъективной оценки из тестового набора случайным образом были выбраны 30 предложений. Мы провели три теста: средний балл мнения (MOS), деградация среднего балла мнения. (DMOS) и предпочтение AB для оценки качества синтезированной речи. Мы набрали 20 носителей мандаринского языка и 10 иностранных студентов-носителей дунганского языка (которые понимали китайцы) в качестве участников. Эти участники прошли обучение перед официальной оценкой. Участники китайского языка оценили акустические модели MSD и MDSM на китайском языке, тогда как участники-дунганы оценивали дунганские акустические модели DSD и MDSD. В ходе теста MOS участники оценили естественность синтезированной речи на 5 баллов. шкала. Представлены средние баллы MOS для синтезированной дунганской и мандаринской речи. на рисунках 7 и 8.

В тесте DMOS синтезированное высказывание каждой модели и соответствующий оригинал запись состояла из пары речевых файлов. Эти пары случайным образом разыгрывались до предметы, причем синтезированная речь предшествует оригиналу. Перед участниками была поставлена задача со скрупулезным сравнением двух файлов и оценкой сходства синтезированных речь оригиналу по 5-бальной шкале. Оценка 5 указывает на то, что синтезированный речь была похожа на оригинал, тогда как балл 1 означал значительное несоответствие. На рисунках 9 и 10 показаны средние показатели DMOS для синтезированных дунганского и мандаринского языков. речь соответственно.

Machine Translated by Google



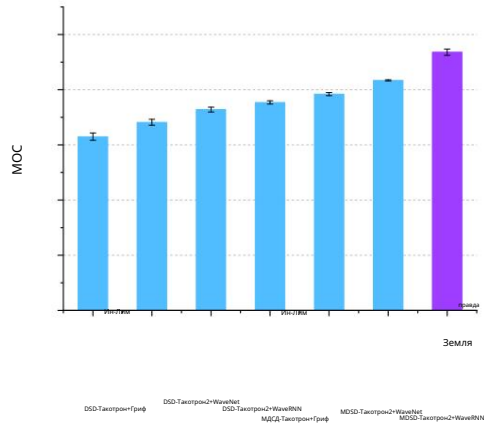


Рисунок 7. Средние баллы MOS синтезированной дунганской речи при доверительном интервале 95%.

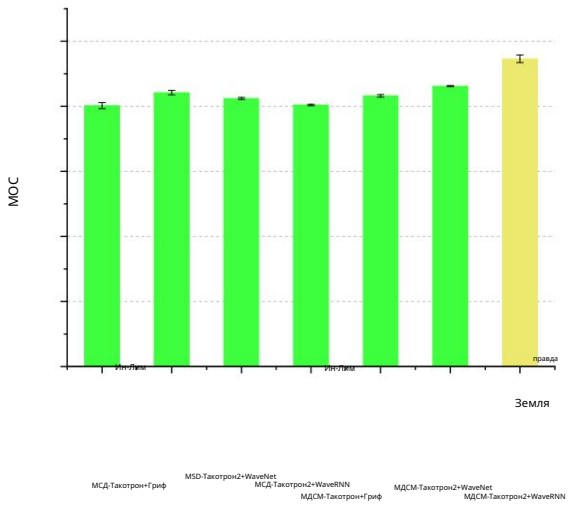


Рисунок 8. Средние баллы MOS синтезированной мандаринской речи при доверительном интервале 95%.

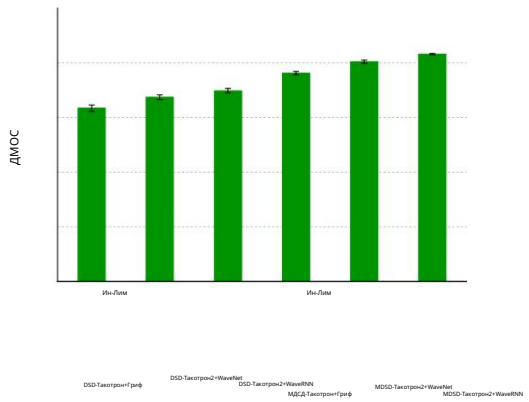


Рисунок 9. Средние баллы DMOS синтезированной дунганской речи при доверительном интервале 95%.

Machine Translated by Google



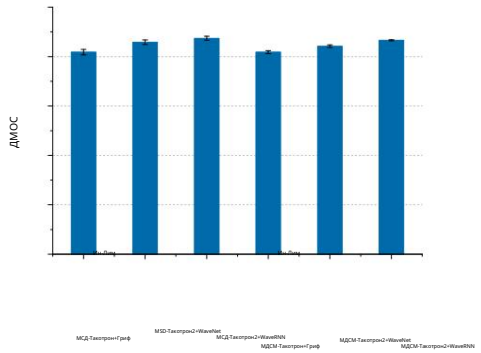


Рисунок 10. Средние баллы DMOS синтезированной мандаринской речи при доверительном интервале 95%.

В тесте предпочтений AB каждая пара состояла из двух одинаковых предложений. Синтезированные высказывания проигрывались в случайном порядке. Участникам было предложено послуш и оцените, какое высказывание имело лучшее качество, или укажите «нейтральное», если нет предпочтений было различено. Синтезированные результаты предпочтений дунганской и мандаринской речи: представлены в таблицах 10 и таблицах 11 соответственно.

Таблица 10. Субъективная оценка предпочтения АБ (%) дунганцев с ρ < 0,01.

| | DSD-Такотрон+ Гриффин-Лим | DSD-Такотрон2+ ВейвНет | DSD-Такотрон2 +WaveRNN | МДСД-Такотрон+ Гриффин-Лим | МДСД-Такотрон2 +ВейвНет | MDSD-Такотрон2+ ВолнаRNN | Нейтральный |
|---|---|---|---|---|---|---|---|
| 1 | 12,7 | 22,9 | 52,6 | - | - | - | 11,8 |
| 2 | 29,5 | 32,0 | 27,6 | - | - | - | 10,9 |
| 3 | - | - | - | 17,7 | - | 69,9 | 12.4 |
| 4 | - | - | - | 3,2 | | 70,8 | 11.3 |
| 5 | - | - | - | - | 17.1 | 72,1 | 10,8 |

Таблица 11. Субъективная оценка предпочтения АВ (%) мандаринского языка с ρ < 0,01.

| | МСД-Такотрон+ Гриффин-Лим | МСД-Такотрон2+ ВейвНет | МСД-Такотрон2+ ВолнаRNN | МДСМ-Такотрон+ Гриффин-Лим | МДСМ-Такотрон2+ ВейвНет | МДСМ-Такотрон2+ ВолнаRNN | Нейтральный |
|---|---|---|---|---|---|---|---|
| 1 | - | 24,54 | 63,56 | - | - | - | 11,9 |
| 2 | - | 19,98 | 67,42 | - | - | - | 12,6 |
| 3 | - | - | - | - | 11,8 | 71,9 | 16.3 |
| 4 | - | - | - | 14,4 | - | 75,1 | 10,5 |
| 5 | - | - | - | - | 10,7 | 79,6 | 9,7 |

## 4. Дискуссия

По объективным оценкам, хотя система TTS на основе Tacotron+Griffin-Lim Сопоставляя лингвистические особенности с акустическими покадрово через одноязычный дунганский корпус, синтезированная дунганская речь нуждается в улучшении ее качества и читаемости. Однако направленное внимание и точно настроенная акустическая модель могут улучшить читаемость. и сократить время обучения. Следовательно, Tacotron2+WaveRNN на основе трансферного обучения Акустическая модель Framework превосходит другие. Объективные результаты акустической модели MDSD превосходят результаты акустической модели DSD. Это потому, что дунган – это разновидность северо-западного диалекта Китая, который имеет много внутренних сходств. Учитывая сходство произношения мандаринского и дунганского языков, один и тот же символ обозначает их точное произношение. Таким образом, мы приходим к выводу, что добавление корпуса мандаринского языка и использование Трансферное обучение может улучшить качество и читаемость синтезированной дунганской речи.

Все субъективные оценки совпадают с объективными оценками в различных аспектах. Платформа Tacotron2+waveRNN на основе трансферного обучения обеспечивает превосходное качество речи,

Machine Translated by Google



особенно в отношении естественности и читаемости синтезированной речи. С добавлением корпуса мандаринского языка качество и читаемость синтезированной дунганской речи с использованием фреймворков TTS на основе трансферного обучения превосходят одноязычные фреймворки TTS, обученные на основе корпуса. Это дополнительно подтверждается тестом предпочтений AB, который подтверждает, что предлагаемые нами структуры TTS обеспечивают улучшенное качество и читаемость по сравнению с речью, синтезированной с помощью одноязычной акустической модели.

## 5. Выводы

Это исследование расширяет наши предыдущие исследования, реализуя синтез мандаринской речи на основе трансферного обучения и синтез дунганской речи с низким уровнем ресурсов в рамках Tacotron2 + WaveRNN. Мы также разработали комплексный анализатор дунганского текста. Объективные и субъективные эксперименты показали, что синтез дунганской речи на основе трансферного обучения в рамках платформы Tacotron2+WaveRNN превзошел альтернативные методы и структуру синтеза одноязычной дунганской речи. Более того, трансферное обучение не поставило под угрозу качество речи и читаемость синтезированной дунганской речи с низким уровнем ресурсов. Таким образом, наш подход имеет значительный потенциал для разработки систем синтеза речи для языков меньшинств с низкими ресурсами.

Многочисленные прорывы были достигнуты в TTS на основе глубоких нейронных сетей. Мы заметили, что в последнее время были предложены некоторые новые методы синтеза речи [50–52] . Руководствуясь недавними достижениями в области авторегрессионных (AR) моделей, использующих архитектуры только декодера для генерации текста, в нескольких исследованиях, таких как VALL-E [53] и BASE TTS [54], аналогичные архитектуры применяются к задачам TTS. Эти исследования демонстрируют замечательную способность архитектур, состоящих только из декодеров, воспроизводить естественно звучащую речь. Эти исследования демонстрируют замечательную способность архитектур, состоящих только из декодеров, воспроизводить естественно звучащую речь. Будущие исследования будут сосредоточены на использовании этих новых методов для улучшения качества синтеза речи дунганского языка, уменьшения размера корпуса дунган и достижения синтеза речи для дунганских языков с использованием более крупного корпуса. Кроме того, будет изучено многозадачное обучение для реализации сценариев, независимых от говорящего, и повышения эмоциональности синтезированной дунганской речи.

Вклад авторов: концептуализация, ML и HY; формальный анализ, HY и RJ; курирование данных , ML и RJ; письменное – подготовка первоначального проекта, ML и RJ; написание — обзор и редактирование, HY и ML; надзор, HY; приобретение финансирования, HY. Все авторы прочитали и согласились с опубликованной версией рукописи.

Финансирование: Исследование поддержано исследовательским фондом Национального фонда естественных наук Китая (грант № 62067008).

Заявление Институционального наблюдательного совета: Не применимо к исследованиям без участия людей или животных.

Заявление об информированном согласии: Не применимо.

Заявление о доступности данных: в рукописи мы использовали два набора обучающих данных. Один из них — общедоступный набор данных на китайском языке (THCHS-30), а другой — набор данных на языке донгган, включая речь и текст. Первый является общедоступным и доступен по адресу http://www.openslr.org/18/. (по состоянию на 16 июня 2024 г.). Последний представляет собой набор данных, созданный самостоятельно и не являющийся общедоступным. Однако данные будут предоставлены по запросу.

Конфликты интересов: Авторы заявляют об отсутствии конфликта интересов. Спонсоры не играли никакой роли в разработке исследования; при сборе, анализе или интерпретации данных; в написании рукописи; или в решении опубликовать результаты.

## Рекомендации

1. Ту, Т.; Чен, Ю.Дж.; Чие Йе, К.; Йи Ли, Х. Сквозное преобразование текста в речь для языков с ограниченными ресурсами посредством межъязыкового переноса Обучение. arXiv 2019, arXiv:1904.06508.
2. Лю, Р.; Сисман, Б.; Бао, Ф.; Ян, Дж.; Гао, Г.; Ли, Х. Использование морфологических и фонологических особенностей для улучшения просодической фразировки для синтеза монгольской речи. IEEE/ACM Транс. Аудиоречевой язык. Процесс. 2021, 29, 274–285. [Перекрестная ссылка]
3. Саэки, Т.; Маити, С.; Ли, Х.; Ватанабэ, С.; Такамичи, С.; Саруватари, Х. Текстово-индуктивная языковая адаптация на основе графона для синтеза речи с низкими ресурсами. IEEE/ACM Транс. Аудиоречевой язык. Процесс. 2024, 32, 1829–1844. [Перекрестная ссылка]

Machine Translated by Google



4. Сюй, Дж.; Тан, Х.; Рен, Ю.; Цинь, Т.; Ли, Дж.; Чжао, С.; Лю, Т. Л. Р. Речь: синтез и распознавание речи с крайне низкими ресурсами. В материалах 26-й Международной конференции ACM SIGKDD по обнаружению знаний и интеллектуальному анализу данных, KDD'20, Нью-Йорк, Нью-Йорк, США, 6–10 июля 2020 г.; стр. 2802–2812. [Перекрестная ссылка]

5. Он, М.; Ян, Дж.; Он, Л.; Сунг, Ф.К. Многоязычные модели Byte2Speech для масштабируемого синтеза речи с низким уровнем ресурсов. apXiv 2021, arXiv: 2103.03541.

6. Оливейра, ФС; Казанова, Э.; Младший, АС; Соареш, А.С.; Гальван Фильо, А.Р. CML-TTS: многоязычный набор данных для синтеза речи на языках с ограниченными ресурсами. В тексте, речи и диалоге; Экштейн К., Партл Ф., Конопик М., ред.; Шпрингер: Чам, Швейцария, 2023 г.; стр. 188–199.

7. Чжу, Ю. Дунганский язык: особая разновидность диалектов Шэньси и Ганьсу. Азиатский язык. Культ. 2013, 4, 51–60. 8. Цзян, Ю. Язык дунган и его связь с диалектами Шэньси и Ганьсу. Дж. Чин. Лингвист. 2014, 42, 229–258.

9. Чен, Л.; Ян, Х.; Ван, Х. Исследование синтеза дунганской речи на основе глубокой нейронной сети. В материалах 11-го Международного симпозиума по обработке китайской разговорной речи (ISCSLP) 2018 г., Тайбэй, Тайвань, 26–29 ноября 2018 г.; стр. 46–50. [Перекрестная ссылка]

10. Цзян Р.; Чен, К.; Шан, Х.; Ян, Х. Использование улучшения речи для реализации синтеза речи малоресурсных дунганских языков. В материалах 24-й конференции Восточного Международного комитета COCOSDA по координации и стандартизации речевых баз данных и методов оценки (O-COCOSDA), Сингапур, 18–20 ноября 2021 г.; стр. 193–198. [Перекрестная ссылка]

11. Хант, Эй Джей; Блэк, А.В. Выбор единиц измерения в системе конкатенативного синтеза речи с использованием большой базы данных речи. В материалах Международной конференции IEEE 1996 г. по акустике, речи и обработке сигналов, Атланта, Джорджия, США, 9 мая 1996 г.; Том 1, стр. 373–376.

12. Токуда, К.; Нанкаку, Ю.; Тода, Т.; Дзен, Х.; Ямагиши, Дж.; Оура, К. Синтез речи на основе скрытых марковских моделей. Учеб. IEEE 2013, 101, 1234–1252. [Перекрестная ссылка]

13. Линг, З.Х.; Дэн, Л.; Ю, Д. Моделирование спектральных огибающих с использованием ограниченных машин Больцмана и сетей глубокого доверия для статистического параметрического синтеза речи. IEEE Транс. Аудиоречевой язык. Процесс. 2013, 21, 2129–2139. [Перекрестная ссылка]

14. Дзен, Х.; Старший, А.; Шустер, М. Статистический параметрический синтез речи с использованием глубоких нейронных сетей. В материалах Международной конференции IEEE 2013 г. по акустике, речи и обработке сигналов, Ванкувер, Британская Колумбия, Канада, 26–31 мая 2013 г.; стр. 7962–7966. [Перекрестная ссылка]

15. Ван П.; Цянь, Ю.; Сунг, ФК; Он, Л.; Чжао, Х. Встраивание слов для синтеза TTS на основе рекуррентной нейронной сети. В материалах Международной конференции IEEE по акустике, речи и обработке сигналов (ICASSP) 2015 г., Южный Брисбен, Квинсленд, Австралия, 19–24 апреля 2015 г.; стр. 4879–4883. [Перекрестная ссылка]

16. Ю, К.; Лю, П.; Ву, З.; Анг, СК; Мэн, Х.; Цай, Л. Изучение межъязыковой информации с помощью многоязычного BLSTM для синтеза речи языков с низким уровнем ресурсов. В материалах Международной конференции IEEE по акустике, речи и обработке сигналов (ICASSP) 2016 г., Шанхай, Китай, 20–25 марта 2016 г.; стр. 5545–5549. [Перекрестная ссылка]

17. Тан, Х.; Чен, Дж.; Лю, Х.; Конг, Дж.; Чжан, К.; Лю, Ю.; Ван, Х.; Ленг, Ю.; Йи, Ю.; Он, Л.; и другие. NaturalSpeech: сквозной синтез речи в речь с качеством человеческого уровня. IEEE Транс. Паттерн Анал. Мах. Интел. 2024, 46, 4234–4245. [Перекрестная ссылка]

18. Ван Ю.; Скерри-Райан, Р.Дж.; Стэнтон, Д.; Ву, Ю.; Вайс, Р.Дж.; Джейтли, Н.; Ян, З.; Сяо, Ю.; Чен, З.; Бенджио, С.; и другие. Такотрон: к сквозному синтезу речи. В материалах 18-й ежегодной конференции Международной ассоциации речевой коммуникации, Interspeech 2017, Стокгольм, Швеция, 20–24 августа 2017 г.

19. Шен Дж.; Панг, Р.; Вайс, Р.Дж.; Шустер, М.; Джейтли, Н.; Ян, З.; Чен, З.; Чжан, Ю.; Ван, Ю.; Скеррв-Райан, Р.; и другие. Естественный синтез TTS путем обработки Wavenet на основе предсказаний MEL-спектрограммы. В материалах Международной конференции IEEE по акустике, речи и обработке сигналов (ICASSP) 2018 г., Калгари, АВ, Канада, 15–20 апреля 2018 г.; стр. 4779–4783. [Перекрестная ссылка]

20. Гриффин, Д.; Лим, Дж. Оценка сигнала с помощью модифицированного кратковременного преобразования Фурье. IEEE Транс. Акуст. Речевой сигнальный процесс. 1984, 32, 236–243. [Перекрестная ссылка]

21. ван ден Оорд, А.; Дилеман, С.; Дзен, Х.; Симонян, К.; Виньялс, О.; Грейвс, А.; Кальхбреннер, Н.; Старший, А.; Кавукчуоглу, К. WaveNet: генеративная модель для необработанного аудио. arXiv 2016, arXiv:1609.03499.

22. Кальхбреннер, Н.; Элсен, Э.; Симонян, К.; Нури, С.; Касагранде, Н.; Локхарт, Э.; Стимберг, Ф.; Ван ден Оорд, А.; Дилеман, С.; Кавукчуоглу, К. Эффективный нейронный синтез звука. arXiv 2018, arXiv:1802.08435.

23. Бямбадорж, З.; Нисимура, Р.; Аюш, А.; Охта, К.; Китаока, Н. Система преобразования текста в речь для языка с низкими ресурсами с использованием межъязыкового переноса обучения и увеличения данных. EURASIP J. Аудио-речевая музыка. Процесс. 2021, 2021, 42. [CrossRef]

24. Джоши, Р.; Гарера, Н. Быстрая адаптация говорящего из текста с низкими ресурсами в речевые системы с использованием синтетических данных и трансферного обучения. В материалах 37-й Тихоокеанской азиатской конференции по языку, информации и вычислениям, Гонконг, Китай, 2–4 декабря 2023 г.; Хуанг Ч.Р., Харада Ю., Ким Дж.Б., Чен С., Сюй, Ю.Ю., Черсони, Е.А.П., Цзэн В.Х., Пэн Б., Ли Ю. и др., ред.; ACL: Гонконг, Китай, 2023 г.; стр. 267–273.

25. До, П.; Колер, М.; Дейкстра, Дж.; Клабберс, Э. Стратегии трансферного обучения для синтеза речи с низкими ресурсами: сопоставление телефонов, ввод функций и выбор исходного языка. В материалах 12-го семинара ISCA по синтезу речи (SSW2023), Гренобль, Франция, 26–28 августа 2023 г.; стр. 21–26. [Перекрестная ссылка]

Machine Translated by Google




26. Азиза, К.; Джатмико, В. Трансферное обучение, контроль стиля и потери при реконструкции говорящего для многоязычного многоязычного говорящего с нулевым выстрелом преобразование текста в речь на языках с низкими ресурсами. IEEE Access 2022, 10, 5895–5911. [Перекрестная ссылка]
27. Цай, З.; Ян, Ю.; Ли, М. Межъязыковой синтез речи нескольких говорящих с ограниченными данными двуязычного обучения. Вычислить. Речь Ланг. 2023, 77, 101427. [CrossRef]
28. Ян, Х.; Оура, К.; Ван, Х.; Ган, З.; Токуда, К. Использование адаптивного обучения говорящих для реализации мандаринско-тибетского межъязыкового общения. синтез речи. Мультимед. Инструменты Прил. 2015, 74, 9927–9942. [Перекрестная ссылка]
29. Ван, Л.; Ян, Х. Метод сегментации тибетских слов, основанный на модели bilstm_crf. В материалах Международной конференции IEEE 2018 по обработке азиатских языков (IALP), Бандунг, Индонезия, 15–17 ноября 2018 г.; стр. 297–302.
30. Чжан, В.; Ян, Х.; Бу, Х.; Ван, Л. Глубокое обучение синтезу мандаринско-тибетской межъязыковой речи. IEEE Access 2019, 7, 167884–167894. [Перекрестная ссылка]
31. Чжан, В.; Ян, Х. Улучшение последовательного синтеза тибетской речи с помощью просодической информации. АКМ Транс. Азиатская малоресурсная. Ланг. Инф. Процесс. 2023, 22, 6012. [CrossRef]
32. Чжан, В.; Ян, Х. Метаобучение для синтеза мандаринско-тибетской межъязыковой речи. Прил. наук. 2022, 12, 2185. [CrossRef]
33. Хай Ф. Пилотное исследование заимствованных слов в среднеазиатском дунганском языке. Синьцзянский университет. Дж. 2000, 28, 58–63.
34. Лин Т. Особенности, состояние и тенденции развития тунганского языка в Центральной Азии. Созерцание Лингвист. 2016, 18, 234–243.
35. Глэдни, округ Колумбия. Реляционная инаковость: построение дунганской (хуэйской), уйгурской и казахской идентичностей в Китае, Центральной Азии и Турции. Хист. Антрополь. 1996, 9, 445–477. [Перекрестная ссылка]
36. Мяо, DX двуязычная модель обучения народа дунгань. Дж. Рез. Образование. Этн. Незначительный. 2008, 19, 111–114.
37. Цзя, Ю.; Хуанг, Д.; Лю, В.; Донг, Ю.; Ю, С.; Ван, Х. Нормализация текста в системе преобразования текста в речь на китайском языке. В материалах Международной конференции IEEE 2008 г. по акустике, речи и обработке сигналов, Лас-Вегас, Невада, США, 31 марта – 4 апреля 2008 г.; стр. 4693–4696. [Перекрестная ссылка]
38. Ванмежаси, Н. Исследование нескольких ключевых проблем сегментации тибетских слов. Дж. Чин. Инф. Процесс. 2014, 28, 132–139.
39. Завьялова, О. Дунганский язык. 2015. Доступно онлайн: https://www.academia.edu/42869092/Dungan_Language. (доступ 16 июня 2024 г.).
40. Линь, Т. Письмо Дунган — успешное испытание китайского алфавитного письма. Дж. Во-вторых. Северо-Западный университет. Натл. 2005, 2005, 31–36.
41. Ян, В.Дж.; Чжан Р. Этническая идентичность в межнациональном контексте — пример исследования «дунган» и национальности хуэй. Дж. Саут-Цент. унив. Натл. 2009, 29, 31–36.
42. Чжэн Ю.; Тао, Дж.; Вэнь, З.; Ли, Ю. Сквозное предсказание просодических границ на основе BLSTM-CRF с контекстно-зависимыми встраиваниями во внешнем интерфейсе преобразования текста в речь. Учеб. Интерспич 2018, 9, 47–51. [Перекрестная ссылка]
43. Хлаинг, АМ; Па, WP Модели последовательностей для преобразования графем в фонемы в Большом словаре произношения Мьянмы. В материалах 22-й конференции Восточного Международного комитета COCOSDA по координации и стандартизации речевых баз данных и методов оценки (O-COCOSDA), Себу, Филиппины, 25–27 октября 2019 г.; стр. 1–5. [Перекрестная ссылка]
44. Тан, К.; Сан, Ф.; Конг, Т.; Чжан, В.; Ян, К.; Лю, К. Обзор глубокого трансферного обучения. «Искусственные нейронные сети и машинное обучение» — ICANN 2018; Куркова В., Манолопулос Ю., Хаммер Б., Илиадис Л., Маглогианнис И., ред.; Спрингер: Чам, Швейцария, 2018 г.; стр. 270–279.
45. Ван, Д.; Чжан, Х. THCHS-30: Корпус свободной китайской речи. arXiv 2015, arXiv:1512.01882.
46. Кубичек Р. Мел-кепстральное расстояние для объективной оценки качества речи. В материалах Тихоокеанской конференции IEEE по коммуникационным компьютерам и обработке сигналов, Виктория, Британская Колумбия, Канада, 19–21 мая 1993 г.; Том 1, стр. 125–128. [Перекрестная ссылка]
47. Диман, Дж.К.; Силамантула, К.С. Спектрально-временной метод оценки апериодичности и озвученных/неозвученных границ принятия решений речевых сигналов. В материалах Международной конференции IEEE по акустике, речи и обработке сигналов 2019 г. (ICASSP2019), Брайтон, Великобритания, 12–17 мая 2019 г.; стр. 6510–6514. [Перекрестная ссылка]
48. Кастеласо, И.; Митани, Ю. Об использовании среднеквадратической ошибки в качестве показателя квалификации. Аккредитовать. Квал. Ассур. 2012, 17, 95–97. [Перекрестная ссылка]
49. Рен, Ю.; Тан, Х.; Цинь, Т.; Чжао, С.; Чжао, З.; Лю, Тай. Почти неконтролируемое преобразование текста в речь и автоматическое распознавание речи. arXiv 2020, arXiv:1905.06791.
50. Рен, Ю.; Ху, К.; Тан, Х.; Цинь, Т.; Чжао, С.; Чжао, З.; Лю, Тай. FastSpeech 2: быстрое и качественное сквозное преобразование текста в речь. arXiv 2022, arXiv:2006.04558.
51. Чен Дж.; Песня, Х.; Пэн, З.; Чжан, Б.; Пан, Ф.; Ву, З. LightGrad: Облегченная вероятностная модель диффузии для преобразования текста в речь. В материалах Международной конференции IEEE по акустике, речи и обработке сигналов 2023 г. (ICASSP2023), Остров Родос, Греция, 4–10 июня 2023 г.; стр. 1–5. [Перекрестная ссылка]
52. Го, Ю.; Ду, К.; Ма, З.; Чен, Х.; Ю, К. VoiceFlow: эффективное преобразование текста в речь с исправленным сопоставлением потока. В материалах Международной конференции IEEE 2024 г. по акустике, речи и обработке сигналов (ICASSP2024), Сеул, Республика Корея, 14–19 апреля 2024 г.; стр. 11121–11125. [Перекрестная ссылка]

Machine Translated by Google



53. Ван, К.; Чен, С.; Ву, Ю.; Чжан, З.; Чжоу, Л.; Лю, С.; Чен, З.; Лю, Ю.; Ван, Х.; Ли, Дж.; и другие. Языковые модели нейронных кодеков представляют собой синтезаторы речи с нулевым преобразованием текста. arXiv 2023, arXiv:2301.02111.
54. Лайщак, М.; Камбара, Г.; Ли, Ю.; Бейхан, Ф.; ван Корлаар, А.; Ян, Ф.; Жоли, А.; Мартин-Кортинас, А.; Аббас, А.; Михальский, А.; и другие. BASE TTS: уроки построения модели преобразования текста в речь с миллиардом параметров на основе 100 тысяч часов данных. arXiv 2024, arXiv:2402.08093.

Отказ от ответственности/Примечание издателя: Заявления, мнения и данные, содержащиеся во всех публикациях, принадлежат исключительно отдельному автору(ам) и соавторам(ам), а не MDPI и/или редактору(ам). MDPI и/или редактор(ы) не несут ответственности за любой вред людям или имуществу, возникший в результате любых идей, методов, инструкций или продуктов, упомянутых в контенте.