Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ ИУ «Информатика и системы управления»

КАФЕДРА ИУ-7 «Программное обеспечение ЭВМ и информационные технологии»

РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

НА ТЕМУ:

«Классификация известных методов генерации междометий»

Студент **ИУ7-54Б** _____ **Золотухин А. В.**

Руководитель _____ **Волкова Л. Л.**

2022 г.

РЕФЕРАТ

Научно-исследовательская работа представляет собой анализ существующих методов синтеза речи, а также классификацию существующих методов генерации междометий.

Ключевые слова: синтез речи, междометия, Unit Selection, скрытые марковские модели, модификация записи голоса, TD-PSOLA, SPECINT.

Расчетно-пояснительная записка к научно-исследовательской работе содержит 20 страниц, 1 таблицу, 11 источников.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Синтез речи – искусственное создание звучащей речи человека. Одним из самых важных свойств задачи синтеза является качество получаемой речи. Применение технологии синтеза речи на современном коммерческом уровне зависит именно от качества. Под системами автоматического синтеза речи понимают системы, преобразующие текст и другую информацию в звучащую речь.

Технология автоматического синтеза речи применяется в самых различных отраслях и направлениях:

— телекоммуникации (call-центры);

— мобильные устройства;

— промышленные и бытовые электронные устройства;

— автомобильная индустрия;

— образовательные системы;

— Internet-сервисы;

— системы ограничения доступа;

— аэрокосмическая промышленность;

— военно-промышленный комплекс.

Синтезаторы речи обладают широкими возможностями применения. Например, позвонив в информационную службу, можно уже услышать не роботизированную речь, а приятный естественный голос. Автоинформационная система с технологией синтеза речи, вступит в беседу с каждым дозвонившимся и поможет в получении информации. Такая система освобождает операторов от ответов на часто повторяющиеся вопросы. Также технология синтеза речи открывает широкие возможности для людей с ограниченными возможностями здоровья. Для слепых и слабовидящих разработаны говорящие машины. Для немых предусмотрены специальные устройства синтеза речи, в которых сообщение набирается на клавиатуре, что позволяет им без проблем общаться с другими людьми.

На сегодняшний день благодаря электронным словарям и переводчикам на основе технологии синтеза речи возможно изучение иностранных языков с постановкой правильного произношения. Еще одним примером синтеза речи могут служить различные системы звукового оповещения: телефонная справочная информация, объявление станций в метро, информация об отправлении автобуса или поезда.

Цель научно-исследовательской работы: классификация существующих методов генерации междометий.

Задачами данной научно-исследовательской работы являются:

— описание существующих методов генерации звучащей речи;

— проведение анализа предметной области генерации междометий;

— выделение критериев сравнения методов генерации междометий;

— классификация методов генерации междометий по выделенным критериям.

1 Аналитический раздел

1.1 Методы генерации звучащей речи

С помощью одного метода синтезированная речь создается путем объединения фрагментов записанной речи, хранящихся в базе данных. В другом методе синтезатор моделирует речевой тракт и другие характеристики человеческого голоса для создания полностью синтезированной речи. Качество полученной речи определяется по её сходству с человеческим голосом и по её способности быть понятной.

Система преобразования текста в речь состоит из следующих частей.

1. Графематический анализ — этап, который обеспечивает выделение синтаксических или структурных единиц из входного текста, который может представлять собой линейную структуру, содержащую единый фрагмент текста [1]:

   — выделение в тексте предложений;

   — разметка текста на буквы, цифры, знаки пунктуации и специальные символы;

   — выделение слов в предложениях.

2. Морфологический анализ – переход от словоформ к их леммам (словарным формам лексем), или основам:

   — нормализация текста — расшифровка обозначений и аббревиатур, а также трансформация чисел и знаков в буквенно-словесную форму;

   — определение места ударения и морфо-грамматических характеристик слов в предложении;

   — снятие омонимии.

3. Синтаксический анализ, т.е. выявление грамматической структуры предложений текста — синтаксических связей между словами в предложении.

4. Семантический анализ, при котором определяется смысл фраз, например, в рамках следующих задач:

   — определение тональности текста (общей либо по отношению к некоторому объекту);

   — выделение именованных сущностей (персоналий, локаций, должностей);

   — разрешение анафоры — если считать местоимения указателями, то можно провести аналогию между этой задачей и разыменованием указателя, т.е. требуется определить, на какое слово в тексте ссылается местоимение.

5. Просодическая обработка текста — придание тексту интонационного оформления.

6. Построение транскрипции по правилам.

7. Вычисление физических параметров (частоты основного тона, спектра, интенсивности и длительности) интонации для синтагм синтезируемого текста.

8. Выбор наиболее подходящих звуковых элементов из базы данных (фраз, предложений, слов, слогов и т.д.).

Речевые синтезаторы делятся на два типа по ограниченности словарной базы [2].

1. С ограниченной словарной базой. В синтезаторах с ограниченным словарем речь хранится в виде отдельных слов или предложений, которые выводятся в определенной последовательности в процессе синтеза речевого сообщения. Все фразы в таких системах произносятся диктором заранее.

2. С неограниченной словарной базой. В синтезаторах с неограниченным словарем элементами речи являются фонемы или слоги, поэтому в них слова строятся по фонетическим правилам. Системы данного типа являются перспективными, т.к. они работают с любым подходящим словарем.

На сегодняшний день в сфере синтеза речи выделяют три основные группы методов по принципу работы, а также гибридный метод.

1. Параметрический синтез. Речевой сигнал представлен набором непрерывно изменяющихся во времени параметров. Данный метод речевого синтеза целесообразно использовать в случаях, когда набор текстовых сообщений ограничен и редко подвержен изменению.

2. Компилятивный синтез. Принцип работы данного метода заключается в составлении сообщения из предварительно записанного словаря исходных элементов синтеза. Очевидно, что содержание синтезируемых сообщений фиксируется объёмом словаря.

3. Синтез речи по фонетическим правилам. В этом методе часто в качестве исходных элементов используются полуслоги (дифоны) — сегменты, содержащие половину согласного и половину примыкающего к нему гласного. При этом появляется возможность синтезировать речь по заранее не заданному тексту. При синтезе речи данным методом могут возникать проблемы управления интонационными характеристиками. Качество такого синтеза не совсем соответствует качеству естественной речи, поскольку на границах «сшивки» дифонов часто возникают искажения.

4. Гибридный метод. Оптимальная последовательность звуковых элементов подбирается из речевого корпуса диктора по классическому алгоритму Unit Selection, но с применением статистической интонационной модели, обученной на той же базе, что позволяет повысить естественность звучания синтезируемой речи по сравнению с реализацией на Unit Selection или только на основе технологии скрытых марковских моделях [3].

1.1.1 Оценка качества синтезированной речи

При разработке систем автоматического синтеза речи очень важным является вопрос оценки качества синтеза речи [2]. В процессе оценки качества учитываются следующие основные характеристики:

— разборчивость речи;

— естественность (натуральность) речи;

— мультимодальность речи;

— многоязычие.

Основным критерием оценки качества синтеза речи является разборчивость синтезированной речи. Оценка разборчивости речи производится в соответствии с ГОСТ Р 59880-2021.

Еще одной важной характеристикой, используемой для оценки качества синтезатора речи, является естественность (натуральность) речи. Натуральность речи можно оценить, но нет объективных критериев — это только субъективное впечатление слушателя. Ведь даже разные люди имеют разное произношение, которое иногда может даже показаться неестественным. Такую оценку можно получить с привлечением респондентов.

Мультимодальность речи — это отражение эмоционального состояния говорящего, индивидуальность его голоса, стиль речи, акцент и т.п. — это будут различные модальности речи. В системах автоматического синтеза речи эта характеристика выражается в возможности синтеза различных типов голосов и их индивидуальных особенностей.

Многоязычие относится к лингвистическим способностям и подразумевает возможность синтеза речи на нескольких естественных языках.

1.1.2 Метод Unit Selection

Метод Unit Selection в настоящее время является основной технологией автоматического синтеза речи, так как он позволяет получать синтезированную речь, которая по своим характеристикам наиболее приближена к естественной [4].

Метод является разновидностью конкатенативного синтеза речи, то есть в процессе создания речевого сигнала используются заранее записанные звуки естественной речи. В данном методе для каждой базовой единицы синтеза производится выбор наиболее подходящего элемента из множества вариантов. Для этого записываются специальные звуковые базы,

размер которых может составлять до нескольких десятков часов звучащей речи. В процессе синтеза метод строит оптимальную последовательность звуковых единиц, в которой учитывается, насколько выбранный элемент соответствует описанию необходимых характеристик звука и насколько хорошо каждый из выбранных элементов будет соединяться с соседними. При этом из базы в качестве оптимальных могут быть выбраны не отдельные звуки, а их цепочки или даже целые предложения. Такой подход позволяет минимизировать модификации речевого сигнала, что повышает естественность синтезируемой речи.

Преимущество метода — при наличии подходящих звуков в базе, получается речь высокого качества.

Недостатки метода:

— объем базы данных;

— критична полнота базы звуковых данных, в том числе доя возможности синтеза речи в заданной модальности в базе должны присутствовать фрагменты речи в этой же тональности;

— существует проблема смены диктора;

— качество синтеза ухудшается в случае отсутствия подходящего звукового элемента в базе данных.

1.1.3 Синтез, основанный на скрытых марковских моделях

Данный тип синтеза является гибридом подходов, основанных на правилах и речевом корпусе. В этом случае происходит описание звуковой базы данных параметрической моделью. Параметры обобщаются множеством статистических моделей, представляющих собой скрытые марковские модели, которые содержат в себе шаблоны речевых элементов [5].

По сравнению с методом Unit Selection, подход, в основе которого лежат модели речи, имеет следующие преимущества и недостатки.

Преимущества метода:

— автоматическое обучение моделей с выбором параметров (например, спектральных характеристик, частоты основного тона, длительности и т.д.), которое возможно выполнять на относительно небольшом речевом материале, позволяет существенно сократить объем требуемой памяти;

— в речи не наблюдаются разрывы, присутствующие при конкатенативном синтезе;

— синтез, основанный на моделях, позволяет легко модифицировать характеристики голоса.

Недостатки метода — речь, полученная на основе моделей, более роботизирована, чем при Unit Selection.

1.2 Методы генерации междометий

Междометие — это часть речи, объединяющая неизменяемые слова, которые выражают эмоции и волевые побуждения, не называя их. Проблема синтеза междометий заключается в том, что в разном контексте одно и то же междометие может звучать по-разному из-за разной эмоциональной окраски фразы. Использование междометий — это один из типичных для человеческого общения способов выражения отношения к предмету разговора в повседневной жизни.

Существуют несколько задач генерации междометий.

1. Генерация междометия в текстовом виде, а также выбор интонации и транскрипции.

2. Генерация междометия в звуковом виде на основе заранее заданного текста и, возможно, метаописания междометия.

3. Генерация звучащего междометия на основе модификации готовой речи:

    — аналитическая модификация на основании выбранного математического аппарата и анализа амплитудно-частотных характеристик записи;

— модификация записи на основе машинного обучения.

Результаты решения первой задачи могут служить входом для методов решения второй и третьей задач. Пусть есть запись междометия «ааа» с монотонной интонацией и продолжительностью 2 секунды. Если задать метаописание междометия (например, номерной тип интонации «радость 2»), возможно преобразовать базовую запись согласно требуемой интонации при совпадении транскрипции.

Для русского языка Yandex.SpeechKit — одно из лучших открытых программных средств озвучки речи, но его недостаток заключается в том, что не может генерировать междометия. Также существует сервис от Сбера SaluteSpeech. В отличие от SpeechKit в нем присутствует возможность добавления междометий в синтезируемую речь, но их количество довольно ограничено, а также их настройка производится автоматически [6].

Основные этапы метода синтеза звучащих междометий, который сможет решить третью задачу, следующие.

1. Анализ базовой записи голоса и извлечение признаков записи голоса(например, высоты, частоты, длительности звука).

2. Модификация признаков записи голоса.

3. Синтез модифицированной записи голоса.

Одним из наиболее трудоемких этапов синтеза междометий является модификация признаков записи голоса. Характеристики сигнала бывают трёх типов [7]: временные, частотные, энергетические. Признаки извлекаются для формирования набора характеристик, информативно отражающих свойства исходных данных. Это позволяет уменьшить размерность звуковой базы.

1.3 Алгоритмы модификации записи голоса

Естественность сигнала зависит от объема речевой базы, которая содержит различные звуковые единицы с различной частотой основного тона. В современных системах такие базы достигают десяти часов речи. Од-

нако даже такого количества недостаточно и приходится прибегать к модификации.

1.3.1 Алгоритм TD-PSOLA

Широко распространены алгоритмы, работающие во временной области, наиболее популярным из которых является технология TD-PSOLA (Time-Domain Pitch-Synchronous-Overlap-Add) [8]. Данный алгоритм работает периодосинхронно, т.е. каждый обрабатываемый фрагмент представляет собой один период. Обязательным условием для этого является возможность определить частоту основного тона сигнала с высокой точностью, т.к. от этого напрямую зависит качество работы этого алгоритма. Далее сигнал разбивается на фрагменты, взвешенные окном Хеннинга, которое захватывает два соседних периода с перекрытием в один период.

Эти взвешенные фрагменты затем могут быть перекомбинированы путём перемещения их центров и наложением с добавлением перекрывающихся частей. Непосредственная модификация частоты основного тона выполняется путём распределения полученных взвешенных фреймов на новые значения частоты.

При сохранении длительности фонограммы, в целом слушатели не замечают неестественностей в сигнале при небольших модификациях частоты основного тона [9].

Когда алгоритм применяется для модификации речи, которой выделяются периоды, качество его работы чрезвычайно высоко, и пока степень изменения частоты основного тона не слишком значительна (± 10 %) от оригинала, качество речи может быть «идеальным», в том смысле, что слушатель не может заметить в речи какой-то неестественности. С точки зрения вычислительной нагрузки на аппаратные ресурсы, алгоритм прост и может применяться в приложениях реального времени [10]. Поэтому зачастую TD-PSOLA рассматривается как приемлемое решение для проблемы модификации частоты основного тона. Также, работая во временной области, он вносит неконтролируемые искажения в сигнал и, при уменьшении частоты основного тона, существенно редуцируется энергия на границах «склеек» фреймов.

1.3.2 Алгоритм SPECINT

В связи с психоакустическими эффектами малейшие искажения в относительном положении формант и изменения огибающей (зависимость амплитуды сигнала от времени) основного тона ведут к побочным эффектам, из-за которых речь становится неестественной, непривычной для нашего восприятия, как следствие, человек при её прослушивании быстро утомляется и не может длительное время внимательно её воспринимать. Поэтому одним из основополагающих действий является получение огибающей основного тона исходного сигнала и её воспроизведение на сигнале новой длительности.

Немаловажно сохранение энергетической огибающей (зависимость амплитуды от частоты), поскольку при увеличении или уменьшении частоты основного тона появляются неизбежные её искажения, что также приводит к снижению естественности речи.

Перед тем как понизить или повысить основной тон, увеличить или уменьшить длительность, необходимо получить значения основного тона на всём модифицируемом участке. При модификации нужно изменить требуемые характеристики аллофонов так, чтобы огибающая основного тона осталась прежней, то есть измениться должен только масштаб (частоты и времени), иначе при малейшем изменении спектральной картины будут слышны режущие слух новые интонации в речи даже при незначительных модификациях. На каждом периоде аллофона вычисляется значение его основного тона, заполняется вектор значений. Далее полученная огибающая изменяется по тону, затем путём сплайн-интерполяции она растягивается или сжимается на требуемую длительность. В итоге получается модель аллофона после модификации, под которую модифицируется исходный аллофон.

Модификация сигнала под требуемую модель происходит следующим образом.

Путём дискретного преобразования Фурье (ДПФ) получается спектр сигнала и рассматриваются отдельно вещественные и мнимые его составляющие.

В спектральной области на частотах, кратных частоте периода, обра-

зуются пики(локальные максимумы). Далее эти пики интерполируются на весь диапазон частот, равный половине частоты дискретизации, и вычисляются значения сплайнов в точках, соответствующих пикам нового периода. После выполнения обратного ДПФ получается период с требуемой частотой.

Однако при таком подходе без дополнений невозможно контролировать амплитуду результирующего сигнала, т.е. её абсолютное значение будет отличным от исходного, что сделает сигнал громче или тише [11].

Для сохранения исходных величин амплитуды вычисляется нормирующий коэффициент, на который домножаются значения коэффициентов вещественной и мнимой части. В результате получаются пики, находящиеся на огибающей, которая нормирована таким образом, чтобы после обратного ДПФ получились те же значения амплитуд, как и в исходном сигнале.

Данный алгоритм позволяет получать хорошее качество модификации при увеличении или уменьшении частоты основного тона до двух раз [8]. Особенно хорошие результаты получаются в случаях, когда сигнал уже имеет естественную огибающую частоты основного тона. Хотя для высоких частот основного тона существует лишь малое количество гармоник для точного формирования данной огибающей, что сказывается на качестве результата. Также существенным недостатком для применения данного метода является потребление огромного количества вычислительных ресурсов [8], так как выполняются сложные математические операции, такие как, например ДПФ.

1.4 Классификация алгоритмов модификации записи голоса

Для классификации алгоритмов модификации записи голоса были выбраны следующие критерии.

1. Качество синтезируемой речи.

2. Степень изменения — критерий, показывающий, насколько сильно характеристики исходного звука отличаются от измененных.

3. Побочные эффекты, возникающие после обработки звука алгоритмом (например, при использовании алгоритма TD-PSOLA на границах «склеек» фреймов уменьшается громкость).

4. Трудоемкость алгоритма.

В таблице 1.1 приведена классификация по выделенным критериям.

Таблица 1.1 – Классификация алгоритмов записи голоса

Критерий сравнения	Алгоритм	
	TD-PSOLA	SPECINT
Качество речи	высокое	высокое
Степень изменения	до 10% от оригинала	до 100% от оригинала
Побочные эффекты	редуцируется энергия на границах «склеек» фреймов	изменение длины аллофона
Трудоемкость	низкая	высокая

Из таблицы видно, что алгоритм TD-PSOLA применим в системах когда требуется выполнить быструю но небольшую коррекцию звука. Алгоритм SPECINT хорошо подойдет там, где необходимо довольно сильно изменить голос диктора, при отсутствии временных рамок.

Вывод

В данном разделе были описаны методы синтеза речи, их достоинства и недостатки, а также описаны этапы синтеза междометий. Были проанализированы алгоритмы модификации записи голоса, описаны их достоинства и недостатки.

ЗАКЛЮЧЕНИЕ

Поставленная цель была достигнута: проведена классификация существующих методов генерации междометий.

В ходе выполнения научно-исследовательской работы были решены следующие задачи:

— описаны существующих методов генерации звучащей речи;

— проведен анализ предметной области генерации междометий, сформулированы критерии сравнения методов генерации междометий;

— классифицированы существующие методы генерации междометий.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Автоматическая обработка текстов на естественном языке и компьютерная лингвистика : учеб. пособие / Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. — М.: МИЭМ, 2011. — 272 с.

2. Рыбин С. В. Синтез речи. Учебное пособие по дисциплине «Синтез речи». — СПб: Университет ИТМО, 2014. — 92 с.

3. Чистиков П.Г., Корольков Е.А., Таланов А.О., Соломенник А.И. Гибридная технология синтеза речи на основе скрытых марковских моделей и алгоритма Unit Selection // Приборостроение, 2013, №2. — СПб: ФГБОУ ВПО «СПбНИУ ИТМО». — С. 33 – 38.

4. Фланаган Дж. Анализ, синтез и восприятие речи. — М.: Связь, 1968. — 396 с.

5. Чистиков П.Г. Технология снитеза русской речи на основе крытых марковскх моделей // Научно-технический вестник информационных технологий, механики и оптики, 2012, № 3. — СПб: Университет ИТМО. — С. 149.

6. Разметка синтеза речи SSML. Тег audio [Электронный ресурс]. Режим доступа: `https://developers.sber.ru/docs/ru/va/how-to/conversation/ssml#teg-audio` (дата обращения 09.11.2022).

7. Поддубный М.И., Киреев, К.В., Русин Н.А., Заборских П.С., Елгин Ю.И. Основные параметры и разновидности речеподобных сигналов, применяемых для защиты речевой информации // Состояние и перспективы развития современной науки по направлению «Информационная безопасность». Сборник статей II Всероссийской научно-технической конференции. — Анапа: ФГАУ «Военный инновационный технополис «ЭРА», 2020. — С. 59 – 65.

8. Чистиков П.Г., Рыбин С.В. Проблемы естественности речевого сигнала в системах синтеза // Журнал «Информационные технологии в образовании», 2011, №1. — СПб: СПбГЭТУ «ЛЭТИ». — С. 22 – 30.

9. Taylor Paul. Text-to-Speech Synthesis. Cambridge University Press, 2009. С. 426–433.

10. Mattheyses Wesley, Verhelst Werner, Verhoeve Piet. Robust pitch marking for prosodic modification of spech using TD-PSOLA. 2006. 01. С. 43–46.

11. Главатских И.А., Чистиков П.Г. Метод модификации физических параметров речевого сигнала на основе периодосинхронного Фурье-анализа // Труды XXXVII международной филологической конференции. — СПб: СПбГУ, 2009. — С. 47 – 62.