УДК 004.934

АЛГОРИТМЫ ПРЕОБРАЗОВАНИЯ ТЕКСТА В РЕЧЬ НА ОСНОВЕ РАЗЛИЧНЫХ ФОРМ СИНТЕЗА

Хлопенкова А.Ю., Гришунов С.С., Белов Ю.С.

Московский государственный технический университет имени Н.Э. Баумана, филиал, Калуга, e-mail: fn1-kf@mail.ru

В статье рассматриваются различные существующие методы и алгоритмы синтеза речи. Рассматривается каскадный синтезатор формант, состоящий из последовательно включенных полосовых резонаторов. Описаны методы линейного прогнозирования, предназначенные как для систем кодирования речи, так и для синтеза речи. В статье описаны методы конкатенативного синтеза звука, в которых используется большая база данных исходных звуков, сегментированных на единицы, и алгоритм выбора единиц, который находит единицы, которые лучше всего соответствуют звуку или музыкальной фразе, которую нужно синтезировать, называемой целью. Качество синтезаторов систем TTS оценивается по различным аспектам, включая разборчивость, естественность и предпочтительность синтетической речи, а также факторы человеческого восприятия, такие как грамматическая правильность. В статье описывается недостаток обычного метода линейного прогнозирования, заключающийся в содержании антиформантов. Также поднимается проблема правильного анализа просодии и произношения по письменному тексту. Подчеркивается влияние разработок в области компьютерных наук и искусственного интеллекта на алгоритмы синтезирования речи, которые развивались на протяжении многих лет в ответ на последние тенденции и новые возможности сбора и обработки данных.

Ключевые слова: TTS, методы PSOLA, формантный синтез, конкатенативный синтез, синусоидальные модели, методы, основанные на линейном прогнозировании

ALGORITHMS FOR CONVERSION TEXT TO SPEECH BASED ON DIFFERENT FORMS OF SYNTHESIS

Khlopenkova A.Yu., Grishunov S.S., Belov Yu.S.

Bauman Moscow State Technical University, Kaluga branch, Kaluga, e-mail: fn1-kf@mail.ru

Various existing methods and algorithms of speech synthesis are discussed in the article. A cascade formant synthesizer consists of band-pass resonators connected in series is considered. Linear predictive methods, which designed both for speech coding systems and speech synthesis, are described. The article describes concatenative sound synthesis methods, which use a large database of source sounds, segmented into units, and a unit selection algorithm to match best the sound or musical phrase to be synthesised. The quality of TTS system synthesizers is evaluated from different aspects, including intelligibility, naturalness, and preference of the synthetic speech, as well as human perception factors, such as comprehensibility. The article describes the disadvantage of the conventional linear forecasting method, which is the content of antiformants. Also discussed the problem of correct prosody and pronunciation analysis from written text. The influence of developments in Computer Science and AI the approaches to speech synthesis that was evolving through years in response to the recent trends and new possibilities in data collection and processing, is highlighted.

Keywords: TTS, PSOLA, Formant Synthesis, Concatenative Synthesis, Sinusoidal Models, Linear Prediction

Под синтезом речи подразумевается восстановление формы речевого сигнала по его параметрам из текстового представления. Система преобразования текста в речь (TTS) преобразует текст в речь, анализируя и обрабатывая исходные данные с помощью обработки естественного языка (Natural Language Processing (NLP)), затем используя технологию цифровой обработки сигналов (Digital Signal Processing (DSP)) для преобразования этого обработанного текста в синтезированное речевое представление [1]. Большинство из методов синтеза речи представляют собой TTS на уровне предложения, которые могут учитывать информацию временного ряда во всем предложении. Однако необходимо установить инкрементную TTS, которая выполняет синтез в меньших лингвистических единицах, чтобы реализовать синтез с малой задержкой, пригодный для систем одновременного преобразования речи в речь [2]. Рассмотрим некоторые из алгоритмов.

Цель исследования: рассмотреть различные алгоритмы для синтезирования речевого потока. Выявить основные этапы реализации методов. Определить преимущества и недостатки каждого алгоритма.

Формантный синтез

Одним из наиболее широко используемых методов синтеза последнего десятилетия является формантный синтез, который основан на модели речи с фильтром источника. Выделяют две основные структуры: параллельная и каскадная, однако для повышения производительности используют их совместно. Формантный синтез позволяет

генерировать бесконечное количество звуков, что делает его более гибким, например, по сравнению с методом конкатенации [3].

Техника синтеза формант – это методика TTS, основанная на правилах. Он производит речевые сегменты путем создания искусственных сигналов на основе набора определенных правил, имитирующих структуру формант и другие спектральные свойства естественной речи. Синтезированная речь генерируется на основе аддитивного синтеза и акустической модели. Основными параметрами акустической модели являются голос, частота, уровни шума и т.д., которые изменяются во времени. Системы на основе формант могут управлять всеми аспектами выходной речи, генерируя широкий спектр эмоций и голоса разного тона с помощью некоторых техник просодического и интонационного моделирования.

Обычно требуется как минимум три форманта для получения членораздельной речи. Пяти формант достаточно, чтобы получить речь высокого качества. Обычно словообразовательный формант моделируется двухполюсным резонатором, позволяющим задавать как частоту форманта, так и ее полосу пропускания [3]. Формант (Q) определяется следующей формулой:

$$Q = \frac{Resonanse}{0.5 * EQ\,width\,at\,half\ maximun\,Gain}.$$

То есть величина «Q» вычисляется путем деления центральной частоты кривой (в Гц) на полуширину кривой эквалайзера (измеренную при половине максимального усиления). Возьмем, например, кривую эквалайзера с центральной частотой 1 кГц и шириной (при половине максимального усиления) 200 Гц. Таким образом, Q будет 1000/100, что равно 10. Точно так же, если бы центральная частота оставалась равной 1 кГц, а ширина была бы всего 50 Гц, Q было бы 40.

К достоинствам данного метода можно отнести высокую разборчивость синтезированной речи даже на высоких скоростях без акустических глюков, а также адаптивность для встраиваемых систем, где память и мощность микропроцессора ограничены. Среди недостатков выделяют сложность разработки правил, определяющих время появления источника и динамические значения всех параметров фильтра даже для простых слов и низкую естественность получаемого речевого потока.

Конкатенативный синтез

Методы конкатенативного синтеза звука (Concatenative sound synthesis (CSS)) используют большую базу данных исходных звуков, сегментированных на единицы (блоки), и алгоритм выбора единиц, который находит последовательность единиц, которая наилучшим образом соответствует звуку или фразе, которые нужно синтезировать, называемой целью.

Целевое расстояние C^t соответствует восприятию подобия блока u_i базы данных целевому блоку t_τ. Оно задается как сумма p взвешенных функций расстояния отдельных дескрипторов C_k^t как [4]:

$$C^t\left(u_i, t_\tau\right) = \sum_{k=1}^{p} w_k^t C_k^t\left(u_i, t_\tau\right).$$

Чтобы способствовать выбору единиц из того же контекста в базе данных, что и в целевой, контекстное расстояние C^x учитывает скользящий контекст в диапазоне r единиц вокруг текущей единицы с весами w_j, уменьшающимися с расстоянием j:

$$C^x\left(u_i, t_\tau\right) = \sum_{j=-r}^{r} w_j^x C^t\left(u^{i+j}, t^{\tau+j}\right).$$

В основном используется Евклидово расстояние, нормированное на стандартное отклонение, а r равно нулю. Некоторым дескрипторам требуются специализированные функции расстояния. Для символьных дескрипторов, например класса фонем, требуется справочная таблица расстояний.

Расстояние конкатенации C выражает нарушение непрерывности, создаваемое конкатенацией единиц u_i и u_j из базы данных. Он задается взвешенной суммой q функций расстояния конкатенации дескрипторов C_k^c

$$C^c\left(u_i, u_j\right) = \sum_{k=1}^{q} w_k^c C_k^c\left(u_i, u_j\right).$$

Расстояние зависит от типа блока: объединение основного блока допускает разрывы в высоте звука и энергии, а дополнительный блок – нет. Последовательные блоки в базе данных имеют нулевое расстояние конкатенации. Таким образом, если в базе данных присутствует целая фраза, соответствующая цели, она будет выбрана полностью.

Преимуществом данного метода является высокое качество звука с точки зрения разборчивости. Однако такие системы занимают очень много времени, потому что они требуют огромных баз данных и жестко кодируют комбинацию для формирования этих слов и результирующая речь может казаться менее естественной и безэмоциональной.

Методы PSOLA

Метод PSOLA (Pitch Synchronous Overlap Add) изначально был разработан компанией France Telecom (CNET). Данный метод хотя и не является методом синтеза, но предоставляет возможность плавно объединять предварительно записанные речевые образцы и обеспечивать контроль высоты тона и длительности, поэтому он используется в некоторых коммерческих системах синтеза речи, таких как ProVerbe и HADIFIX.

Базовый алгоритм состоит из трех шагов [5]:

1. Этап анализа, на котором исходный речевой сигнал сначала разделяется на отдельные, но часто перекрывающиеся сигналы краткосрочного анализа.

2. Преобразование каждого сигнала анализа в сигнал синтеза.

3. Этап синтеза, на котором эти сегменты рекомбинируются посредством перекрывающегося добавления. Кратковременные сигналы $x_m(n)$ получаются из цифрового речевого сигнала $x(n)$ путем умножения сигнала на последовательность сегментов анализа с синхронизацией по высоте тона $h_m(n)$:

$$x_m(n) = h_m(t_m - n)x(n),$$

где m – индекс кратковременного сигнала,
t_m – последовательные моменты, называемые шагами.

Эти метки устанавливаются с синхронной частотой тона для вокализованных частей сигнала и с постоянной скоростью для невокализованных частей. Используемая длина сегмента пропорциональна локальному периоду основного тона, а коэффициент сегмента обычно составляет от 2 до 4. Маркеры основного сегмента определяются либо вручную, либо автоматически некоторыми методами оценки. Рекомбинация сегментов на этапе синтеза выполняется после определения новой последовательности меток основного тона.

TD-PSOLA имеет следующие недостатки: оптимальная разметка шага не является полностью автоматической, а шаг, фаза и рассогласование спектральных амплитуд не позволяют адекватно сглаживать конкатенации. Более того, он предлагает несколько возможностей сжатия базы данных (скорость хранения 80000 бит/с может быть достигнута с помощью кодировщика DPCM с нулевым отводом). Однако он обеспечивает хорошее качество сегментов, а его вычислительная нагрузка очень низка: 7 операций на образец [6].

Методы, основанные на линейном прогнозировании

Способы линейного прогнозирования изначально разработаны для систем кодирования речи, но могут также использоваться в синтезе речи. Фактически первые синтезаторы речи были разработаны из кодировщиков речи. Как и формантный синтез, базовый метод линейного прогнозирования (Linear Predictive Coding (LPC)) основан на модели речи с фильтром источника. Коэффициенты цифрового фильтра оцениваются автоматически из кадра естественной речи.

Основная идея линейного прогнозирования заключается в возможности аппроксимировать речевую выборку $y(n)$ или предсказать ее из набора предыдущих p выборок от $y(n-1)$ до $y(nk)$ линейной комбинацией с малой погрешностью $e(n)$, называемой остаточным сигналом. Тогда получим следующие выражения [5]:

$$y(n) = e(n) + \sum_{k=1}^{p} a(k)\, y(n-k),$$

$$e(n) = y(n) - \sum_{k=1}^{p} a(k)\, y(n-k) = y(n) - \tilde{y}(n),$$

где $\tilde{y}(n)$ – предсказанное значение,
p – порядок линейного предсказателя,
$a(k)$ – коэффициент линейного предсказания, образуемый путем минимизации суммы квадратов отклонений.

Для вычисления этих коэффициентов обычно используются два метода: метод ковариации и метод автокорреляции, однако только с помощью метода автокорреляции фильтр гарантированно будет стабильным.

На этапе синтеза используемое возбуждение аппроксимируется последовательностью импульсов для звонких звуков и случайным шумом для невокализованных звуков. Затем для обработки полученного сигнала используется цифровой фильтр, коэффициенты которого равны $a(k)$. Порядок фильтра обычно составляет от 10 до 12 при частоте дискретизации 8 кГц, однако если требуется получить более высокое качество с частотой дискретизации 22 кГц, то используется порядок от 20 до 24. Коэффициенты обычно обновляются каждые 5–10 мс.

Главный недостаток обычного метода линейного прогнозирования состоит в том, что он представляет собой многополюсную модель, а это означает, что фонемы, содержащие антиформанты, такие как носовые и назализованные гласные, моделируются плохо. Качество также низкое с короткими взрывными согласными (плозивы, эксплозивные,

чистые смычные), потому что события временного масштаба могут быть короче, чем размер кадра, используемый для анализа. Из-за этих недостатков качество синтеза речи стандартным методом LPC обычно считается плохим, но с некоторыми модификациями и расширениями базовой модели качество может быть улучшено [5, 6].

Искаженное линейное прогнозирование (Warped Linear Prediction (WLP)) использует преимущества человеческого слуха, и необходимый порядок фильтрации затем значительно снижается с порядков 20–24 до 10–14 с частотой дискретизации 22 кГц. Основная идея заключается в том, что единичные задержки в цифровом фильтре заменяются следующими всепроходными участками:

$$\tilde{z}^{-1} = D_1(z) = \frac{z^{-1} - \lambda}{1 - \lambda z^{-1}},$$

где λ – параметр деформации между -1 и 1, $D_1(z)$ – элемент задержки деформации, а со шкалой Барка это $l = 0{,}63$ с частотой дискретизации 22 кГц.

WLP обеспечивает лучшее качество на низких частотах и хуже работает на высоких частотах.

Синусоидальные модели

Синусоидальные модели базируются на теории, что речь может быть представлена в виде суммы синусоидальных волн с амплитудами и частотами, которые изменяются во времени. В базовой модели речевой поток s(n) представляется в виде суммы небольшого количества L синусоид [5]:

$$s(n) = \sum_{l=1}^{L} A_l \cos(\omega_l n + \varphi_l),$$

где $A_l(n)$ и $\varphi_l(n)$ представляют собой амплитуду и фазу синусоидальных составляющих, связанных с частотной дорожкой ω_l. Для поиска параметров $A_l(n)$ и $\varphi_l(n)$, вычисляется дискретное преобразование Фурье сегментных сигнальных кадров, и для каждого кадра выбираются экстремумы спектральной величины. Базовая модель, известная как модель Маколея/Кватиери, имеет некоторые модификации, такие как модели ABS/OLA (анализ путем синтеза/добавления перекрытия) и модели гибридного/синусоидального шума.

Синусоидальные модели хорошо подходят для представления гласных и звонких согласных, но плохо работают для представления глухих звуков.

Заключение

Синтез текста в речь – это быстро развивающийся аспект компьютерных технологий, который играет все более важную роль в том, как мы взаимодействуем с системой и интерфейсами на различных платформах. Формантный и конкатенативный методы наиболее часто используются в современных системах синтеза речи. На протяжении длительного времени преобладал формантный синтез, но сегодня все более популярным становится метод конкатенации. В настоящее время существует несколько направлений, основанных на модернизации конкатенативного синтеза текста в речь. Одним из основных направлений является инвентаризация логических единиц.

До сих пор из-за соображений вычислительной сложности и требований к памяти в хранилищах размещалось не более пары токенов для каждой необходимой единицы. Далее, в зависимости от контекста, токены изменялись по таким параметрам, как высота амплитуда и длительность. Основная задача заключалась в оптимальном подборе токенов.

Не так давно стало популярным повышение количества хранимых токенов, части которых могут появляться в различных контекстах, с различным шагом и продолжительностью. Это отрицательно сказывается на работе, так как требуется гораздо больше объема памяти и большего объема вычислений для поиска в большом инвентаре. Однако благодаря развитию современных процессоров доступная память и вычислительная скорость расширяются очень быстро и становятся вполне доступными [7].

Список литературы

1. Хлопенкова А.Ю., Белов Ю.С. Методы обработки естественного языка в виртуальных голосовых помощниках // Электронное периодическое издание E-Scio. 2019. [Электронный ресурс]. URL: http://e-scio.ru/wp-content/uploads/2019/11/Хлопенкова А.-Ю.-Белов-Ю.-С..pdf (дата обращения: 11.09.2021).

2. Takaaki Saeki, Shinnosuke Takamichi, Hiroshi Saruwatari. Incremental Text-to-Speech Synthesis Using Pseudo Lookahead with Large Pretrained Language Model. IEEE Signal Processing Letters. 2020. [Electronic resource]. URL: https://arxiv.org/pdf/2012.12612.pdf (date of access: 11.09.2021).

3. Sudoh K., Kano T., Novitasari S., Yanagita T., Sakti S., Nakamura S. Simultaneous speech-to-speech translation system with neural incremental ASR, MT, and TTS. Vol. abs/2011.04845. 2020. [Electronic resource]. URL: https://arxiv.org/abs/2011.04845 (date of access: 11.09.2021).

4. Sciforce. Text-to-Speech Synthesis: an Overview. Sciforce. 2020. [Electronic resource]. URL: https://medium.com/sciforce/text-to-speech-synthesis-an-overview-641c18fcd35f (date of access: 11.09.2021).

5. Kuligowska K., Kisielewicz P. and Wlodarz A. Speech synthesis systems: disadvantages and limitations. International Journal of Engineering & Technology. 2018. [S.l.]. V. 7. n. 2.28. P. 234–239.

6. Simon Robinson, Gary Marsden and Matt Jones. There's Not an App for That. Morgan Kaufmann. 2015. DOI: 10.1016/C2013-0-00032-1.

7. Хлопенкова А.Ю., Белов Ю.С. Подходы компьютерной лингвистики в технологиях работы голосовых помощников // Всероссийская научно-техническая конференция, г. Калуга, КФ МГТУ им. Н.Э. Баумана. 2019. Т. 3. С. 186–188. [Электронный ресурс]. URL: http://conference.bmstu-kaluga.ru/uploads/userfiles/tom3_sek_10_16.pdf (дата обращения: 11.09.2021).