Оценка качества синтезированной моделями Text-to-Speech речи

И. А. Майбородина
Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина)
iamaiborodina@etu.ru

***Аннотация*. В данной работе выделяются ключевые характеристики синтезированной современными нейронными сетями речи, требующие оценки. Анализируются основные подходы к оценке: как субъективные (основанные на экспертной оценке и восприятии человека), так и объективные метрики, а также современные методы их автоматизации, способные минимизировать необходимость привлечения человеческих экспертов при сохранении достоверности результатов. Это стандартизует процесс оценки и приведёт к ускорению разработки TTS-моделей и повышению их надёжности, сокращению бюджета проектов, улучшению пользовательского опыта.**

Ключевые слова: нейронные сети, машинное обучение, синтез речи, TTS, оценка качества, субъективные метрики, объективные метрики, автоматизация оценки

I. Введение

В настоящее время глубокое обучение обрело большую популярность во всех сферах человеческой деятельности. Это коснулось, в том числе, и области обработки и генерации речи. Если ранее синтез речи основывался на конкатенативных методах, использующих заранее записанные фрагменты речи, или статистических подходах, таких как скрытые марковские модели (HMM), то сейчас существует множество инструментов, созданных на основе нейронных сетей.

Однако современные методы разработки архитектур TTS-систем не гарантируют идеальные результаты и не избавляют от трудностей, связанных с оценкой качества синтезированной речи. Важно найти оптимальные подходы для разных случаев, когда требуется оценить пригодность модели TTS для конкретной задачи.

Качество и естественность речи имеют критическое значение во многих современных приложениях. Ошибки в интонации и артикуляции снижают доверие пользователей к голосовым помощникам, таким, как Google Ассистент, сокращая частоту взаимодействия с системой. Аудиокниги с улучшенной выразительностью синтеза речи получают больше положительных отзывов – например, Audible от Amazon, а неестественность синтезированной речи в медицинских приложениях вызывает у пользователей дискомфорт.

Таким образом, создание универсальных, гибких и автоматизированных методов оценки моделей синтеза речи остается актуальной научно-практической задачей. В данной работе рассматриваются различные оценочные метрики, особое внимание уделяется способам автоматизации оценки, что ускоряет тестирование моделей и минимизирует человеческий фактор.

II. Классификация речевых характеристик

Прежде чем рассматривать конкретные методы оценки качества синтезированной речи, необходимо определить её характеристики. Их немало, и они могут быть разделены на несколько категорий: лингвистические, просодические, акустические (звуковые), по сходству с целевым голосом (для систем, поддерживающих клонирование голоса), адаптивные.

К лингвистическим можно отнести правильность произношения слов (фонетическую точность) и акцентологию (корректную постановку ударений в словах). Правильные ударения улучшают восприятие смысла слушателями.

Просодические характеристики – это соответствие интонационных конструкций смысловой нагрузке и знакам препинания, темп и ритм, выразительность (которая, по сути, определяется сочетанием интенсивности, тональности, громкости и длительности звуков), динамический диапазон (естественные перепады громкости), грамотная расстановка пауз и их продолжительность, эмоциональная окраска. В некоторых моделях TTS, например, Yandex SpeechKit или Silero, существует возможность разметки текста, позволяющей вручную расставить ударения в словах и задать продолжительность пауз. В модели Matcha-TTS существует возможность контролировать темп произношения, и результат звучит естественно, потому что это осуществляется не простым ускорением или замедлением аудио, а изменением скорости произношения каждого слова и регуляцией пауз между словами. Но в большинстве моделей TTS такая возможность отсутствует или, по крайней мере, является неочевидной и выявляется экспериментальным путём. Поэтому в ряде случаев проблемы, связанные с лингвистическими и просодическими характеристиками синтезированной речи, решаются только путём файн-тюнинга на множестве примеров, демонстрирующих контекст. Аналогичным образом можно усилить адаптивные возможности модели, то есть её гибкость, а именно языковую (качество синтеза для разных языков и акцентов) и контекстную (корректная обработка омографов и многозначных конструкций) адаптации, а также робастность (устойчивость к влиянию опечаток и верное произношение необычных примеров: например, сокращений или числительных, написанных цифрами).

К акустическим характеристикам относятся чистота сигнала (отсутствие посторонних шумов, артефактов, спектральных искажений, приводящих к «искусственности» звучания), плавность, стабильность голоса и артикуляционная чёткость (разборчивость произношения звуков). Наконец, если важно сходство с

голосом, на котором модель обучалась (или, в случае с zero-shot моделями – с голосовым промптом), рассматриваются такие особенности синтезированной речи, как приближенность тембра к оригиналу и передача характерных речевых особенностей. Надо заметить, что идеального сходства нельзя добиться постобработкой с помощью моделей непосредственно для клонирования голоса, так как это не сохранит индивидуальные особенности говорящего.

Также в качестве дополнительного критерия для оценки TTS-систем можно рассматривать их вычислительную эффективность.

Естественность речи, или её сходство с человеческой, является совокупностью, так или иначе, всех вышеприведённых категорий характеристик, так как все они оказывают на неё влияние. Даже в какой-то мере скорость генерации (люди не думают минутами перед кратким ответом), что относится к вычислительной эффективности. Ошибки при произношении слов, неправильные ударения выдают искусственную природу речи даже при идеальной акустике, как и монотонность, не говоря о том, что голос со спектральными искажениями наталкивает на мысли о роботах. Низкая адаптивность к контексту или языку нарушает целостность восприятия. При дубляже отсутствие сходства голоса с оригинальным может заставить зрителя ощутить дискомфорт, даже если речь технически корректна и в целом звучит естественно.

Однако не только естественность, но и некоторые другие характеристики тесно взаимосвязаны с другими. Фонетическая точность и акцентология влияют на разборчивость. Последнее также важно для просодики, как и плавность, стабильность голоса. Гибкость, контекстная и языковая адаптивность имеют значение для сохранения индивидуальных черт говорящего.

Более подробно взаимосвязь характеристик можно рассмотреть на рис. 1.

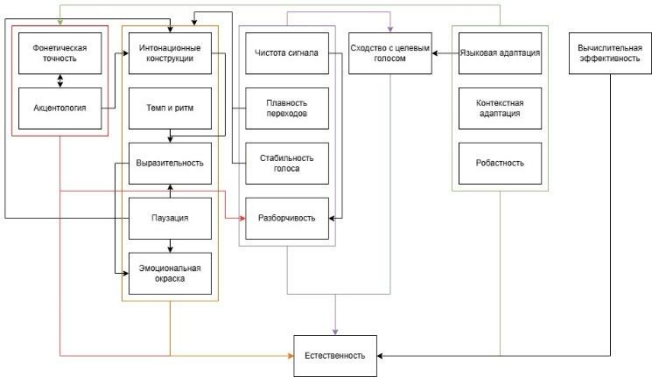

Рис. 1. Взаимосвязь характеристик

В зависимости от задачи наиболее важными могут являться различные сочетания этих характеристик. К примеру, для озвучивания аудиокниг на первый план выходят выразительность и гибкость, а для дубляжа фильмов при переводе на другие языки имеет значение всё вышеуказанное, включая даже сходство голоса с оригиналом. Тем не менее, в любом случае, когда дело касается синтеза речи, а не, к примеру, систем связи, следует оценивать совокупность этих характеристик.

III. Метрики: субъективные и объективные

Перейдём, наконец, к оценке параметров синтезированной речи. Главным субъективным показателем естественности и качества звука является MOS (Mean Opinion Score). Также существуют такие его разновидности, как DMOS и CMOS. Он подразумевает под собой оценку синтезированной речи людьми-экспертами в диапазоне от 1 до 5. Очевидные недостатки данной метрики в том, что, во-первых, оценка субъективна, т. е. основана на восприятии отдельных людей, во-вторых, требуется большое количество респондентов, и в-третьих, это означает усложнение оценки и зависимость от участия в её процессе людей.

В целом, любая из перечисленных характеристик может быть оценена экспертом с помощью дополнительной субъективной метрики. Из их недостатков не следует то, что необходимо полностью исключить такие методы оценивания. Объективные метрики не всегда коррелируют с человеческим восприятием. Синтезированный аудиоконтент создаётся именно для удобства человека, поэтому только люди способны в полной мере оценить естественность звучания искусственной речи.

Под объективными метриками понимают такие метрики, с помощью которых возможно оценить качество без непосредственного участия людей в оценке. Как правило, они основываются на математических формулах, расчёте соотношений показателей звука. Более ранние объективные метрики часто подразумевают сравнение с эталоном, поздние же, появившиеся в ходе развития машинного обучения, могут и не требовать образцовой записи. В отличие от субъективных методов, они обеспечивают воспроизводимые и стандартизированные измерения, что особенно полезно для контроля качества TTS-систем в процессе разработки.

Одними из самых известных и простых метрик для систем распознавания речи являются WER (Word Error Rate) и CER (Character Error Rate). Их можно применить и для оценки синтезированной речи, используя качественную, с низким показателем ошибок, ASR-модель (например, Whisper) и сравнивая полученную транскрипцию с исходным текстом. Это покажет, насколько точно модель TTS воспроизводит отдельные слова и фонемы. Дополнительно можно вычислить процент правильно расставленных в словах ударений, но в этом ASR-система не поможет.

Для оценки естественности интонации синтезированной речи в некоторых работах [1] применяется Pitch RMSE, или F0 RMSE – среднеквадратичная ошибка, демонстрирующая, насколько основной тон синтезированной речи отклоняется от эталонного, или насколько адекватна воспроизведённая моделью интонация для данного контекста. Также в [1] упоминается использование NLL для оценки правдоподобности тона относительно истинного, однако метрика признаётся ненадёжной, так как может быть искусственно улучшена. Причём в приведённой работе сделан вывод о том, что обе метрики слабо коррелируют с MOS. Впрочем, оценка естественности интонации может основываться на анализе интонационных конструкций. Так, например, в русском языке их выделяют 7 – ИК-1, ИК-2 и т. д.

Для оценки скорости речи обычно применяют такие простые метрики, как «количество слов в минуту» (WPM) или «количество слогов в секунду» (SPS). Дополнительно можно выделить Syllable Duration Variability (показатель равномерности произносимых слогов по длительности). Также работы [2] и [3] указывают на широкое применение метрики Pairwise Variability Index для сравнения соседних элементов речевых фрагментов для количественной оценки их ритмических особенностей.

Для измерения такой составляющей выразительности, как громкость, существует Dynamic Range – разница между самым громким и самым тихим участком аудиосигнала.

Одним из подходов к прогнозированию эмоций, присутствующих в речевом фрагменте, является алгоритмическое измерение трёх составляющих, на которые можно разделить каждую эмоцию, представив её вектором в трёхмерном пространстве. Это VAD, или Valence-Arousal-Dominance [4]. Под «valence» понимают позитивность/негативность переживания, «arousal» – энергичность человека, испытывающего эмоцию, «dominance» – степень чувства контроля над ситуацией (доминирование и подчинение). Используя эти шкалы, можно оценить правильность эмоции в синтезированной речи.

Базовой акустической метрикой является SNR – отношение сигнал/шум, демонстрирующее, как понятно из названия, насколько речевой сигнал искажён фоновым шумом. Старыми психоакустическими, т. е. основанными на восприятии звука человеком, методами являются WSS и PESQ. Обе метрики требуют эталонной записи, с которой должно производиться сравнение синтезированной речи. WSS измеряет искажения спектра в 25 критических полосах слуха. PESQ также оценивает искажения спектра, наличие артефактов и шумов, временные задержки и прерывания и выдаёт оценку качества сигнала, коррелирующую с MOS. PESQ признана стандартом [5]. В основном, перечисленные методы оценки акустических характеристик давно используются для тестирования систем связи.

Для оценки стабильности голоса, а также выявления монотонности (при низких показателях) используют jitter и shimmer, измерения «дрожания» основного тона и громкости. Стандартизированным инструментом для оценки разборчивости речи является STOI (Short-Time Objective Intelligibility). Он сравнивает искажённые аудиозаписи с эталоном и выдаёт значение, показывающее, насколько речь понятна для человека, и показатель коррелирует с MOS. Причём, в отличие от SII и STI, STOI учитывает нелинейные искажения.

Более современной метрикой является расстояние Фреше (FAD) [6], которое используется для алгоритмов улучшения аудио. В отличие от остальных акустических показателей, она не требует наличия обязательного эталона. Вместо этого статистика эмбеддингов тестового набора сравнивается с заранее вычисленной статистикой эмбеддингов обучающего набора качественных аудио. Эмбеддинги генерируются с помощью модели – например, VGGish.

Для оценки сходства синтезированного голоса с целевым при клонировании голоса, например, в [7] и других работах используется метрика Speaker Embedding Cosine Similarity, которая основана на сравнении угла между векторами эмбеддингов спикеров.

Таким образом, здесь перечислены наиболее распространённые субъективные и объективные метрики для оценки качества разных характеристик речи – в контексте данной работы – синтезированной. В идеале следует комбинировать объективные метрики с субъективными для точного заключения о качестве речи, что и практикуется во многих существующих исследованиях.

IV. Автоматизация оценки качества синтезированной речи

Автоматизация метрик важна при обучении и валидации TTS-моделей, их тестирования для значительного ускорения процесса проверки, избавления от необходимости привлечения к оценке экспертов и проведения кропотливых ручных расчётов.

Вычисление объективных показателей автоматизировать проще, потому что зачастую для них существуют алгоритмы и формулы. Прогнозирование субъективной оценки стало возможно, только благодаря появлению машинного обучения и, в частности, нейронных сетей. Например, существует такой инструмент на основе глубокого обучения, предсказывающий мнение слушателей-экспертов, как MOSNet.

Для некоторых метрик реализованы отдельные библиотеки, которые уже содержат функции для вычисления различных показателей. Так, jiwer – это пакет Python, позволяющий производить расчёт WER и CER. Для PESQ давно существует официальный инструмент от ITU, написанный на языке C, но также была создана и оболочка для пользователей Python (pesq). STOI можно рассчитать с помощью пакета Python pystoi.

Часть показателей не имеет готовых реализаций, но в теории для их автоматизации могут быть использованы отдельные специальные пакеты. Например, SPS, SDV и PVI можно реализовать с помощью pyphen (модуль Python для переноса по слогам), aeneas (библиотека для синхронизации текста с аудио) и ASR-модели в случае, если нет заготовленной расшифровки речи.

Стоит упомянуть универсальную библиотеку librosa для низкоуровневого анализа аудио. Она позволяет получить множество данных о записи – от общей продолжительности до мел-спектрограммы, – выполнить их оценку и изменить некоторые характеристики: к примеру, частоту дискретизации. Используя функции из этого пакета, можно реализовать алгоритмы части вышеприведённых метрик, для которых не имеется готовых функций и которые достаточно просты, чтобы не требовать обучения моделей. Например, WPM (разделить количество слов на продолжительность, полученную с помощью данной библиотеки), или получение F0 для подсчёта F0 RMSE, громкости – для Dynamic Range.

Также существует широко используемый инструмент для фонетического анализа речи Praat. С его применением можно рассмотреть интонационный контур речи, определить jitter, shimmer и другое.

Библиотека Parselmouth предоставляет интерфейс Python для прямого обращения к коду Praat на C/C++.

Для более сложных метрик, таких как, например, анализ эмоций на основе VAD, понадобится глубокое обучение. Так, чтобы вычислить Speaker Embedding Cosine Similarity, эмбеддинги говорящих можно извлечь с помощью, например, модели ECAPA-TDNN. Предварительно обученные вариации ECAPA-TDNN можно найти в инструментарии SpeechBrain на Hugging Face.

V. Заключение

Оценка синтезированной речи – нетривиальная задача, требующая комплексного подхода. В данной работе были рассмотрены характеристики речевого сигнала и проанализированы существующие методы их оценки, предложены как субъективные, так и объективные метрики, применяемые для оценки качества синтезированной речи, а также возможные пути их автоматизации.

Развитие методов обработки сигналов и машинного обучения влечёт за собой прогресс в области оценки синтезированной речи. Разрабатываются новые, более современные инструменты, в основе которых лежит искусственный интеллект. Трансформеры, появившаяся в последние годы архитектура глубоких нейронных сетей, открывают новые перспективы для создания эффективных инструментов оценивания речевого сигнала за счёт возможности обработки длинных последовательностей и лучшего учёта контекста (что позволит лучше оценивать интонацию), благодаря механизму self-attention, а также лучшей адаптации к языкам. То, что использование трансформеров даёт преимущество по сравнению с более старыми подходами, доказали авторы статьи [10].

Ключевым этапом в оценке TTS-систем является выбор подходящих метрик и оптимального набора инструментов для автоматизации их расчёта для повышения эффективности. Грамотное сочетание субъективных и объективных методов оценки позволяет не только сравнивать разные модели между собой, упрощая и стандартизируя этот процесс, но и контролировать качество синтеза на этапе обучения.

Дальнейшее развитие и совершенствование способов оценки синтезированной речи способствует не только улучшению самих TTS-моделей, но и упрощает процесс их внедрения в реальные приложения. Использование современных метрик и инструментов их автоматизации открывает новые возможности для создания более естественных и качественных систем преобразования текста в речь.

Благодарность

Выражаю благодарность своему научному руководителю Тимофееву Александру Викторовичу за ценные советы при планировании исследования и рекомендации по оформлению статьи.

Список литературы

[1] Controllable Neural Prosody Synthesis. – URL: https://arxiv.org/pdf/2008.03388 (дата обращения: 08.04.2025).

[2] The Pairwise Variability Index and Coexisting Rhythms in Language. [Электронный ресурс]. 2009. – URL: https://www.researchgate.net/publication/24355703_The_Pairwise_Variability_Index_and_Coexisting_Rhythms_in_Language (дата обращения: 11.04.2025).

[3] F0-based Pairwise Variability Index: A Prosodic Metric for Holistic Language Processing. – URL: https://www.internationalphoneticassociation.org/icphs-proceedings/ICPhS2023/full_papers/412.pdf (дата обращения: 11.04.2025).

[4] Affect representation and recognition in 3D continuous valence–arousal–dominance space. – URL: https://link.springer.com/content/pdf/10.1007/s11042-015-3119-y.pdf (дата обращения: 12.04.2025).

[5] Perceptual Evaluation of Speech Quality (PESQ), the new ITU standard for end-to-end speech quality assessment. Part II – Psychoacoustic model. – URL: http://47.115.32.177/media/upfile/PESQ_20230620014449_469.pdf (дата обращения: 13.04.2025).

[6] Fréchet Audio Distance: A Metric for Evaluating Music Enhancement Algorithms. – URL: https://arxiv.org/pdf/1812.08466 (дата обращения: 14.04.2025).

[7] MetricCycleGAN-VC: Forcing CycleGAN-Based Voice Conversion Systems to Associate With Objective Quality Metrics. 2023. – URL: https://ieeexplore.ieee.org/abstract/document/10701302 (дата обращения: 14.04.2025).

[8] Transformer Networks for Non-Intrusive Speech Quality Prediction. – URL: https://www.isca-archive.org/interspeech_2022/jayesh22_interspeech.pdf (дата обращения: 15.04.2025).