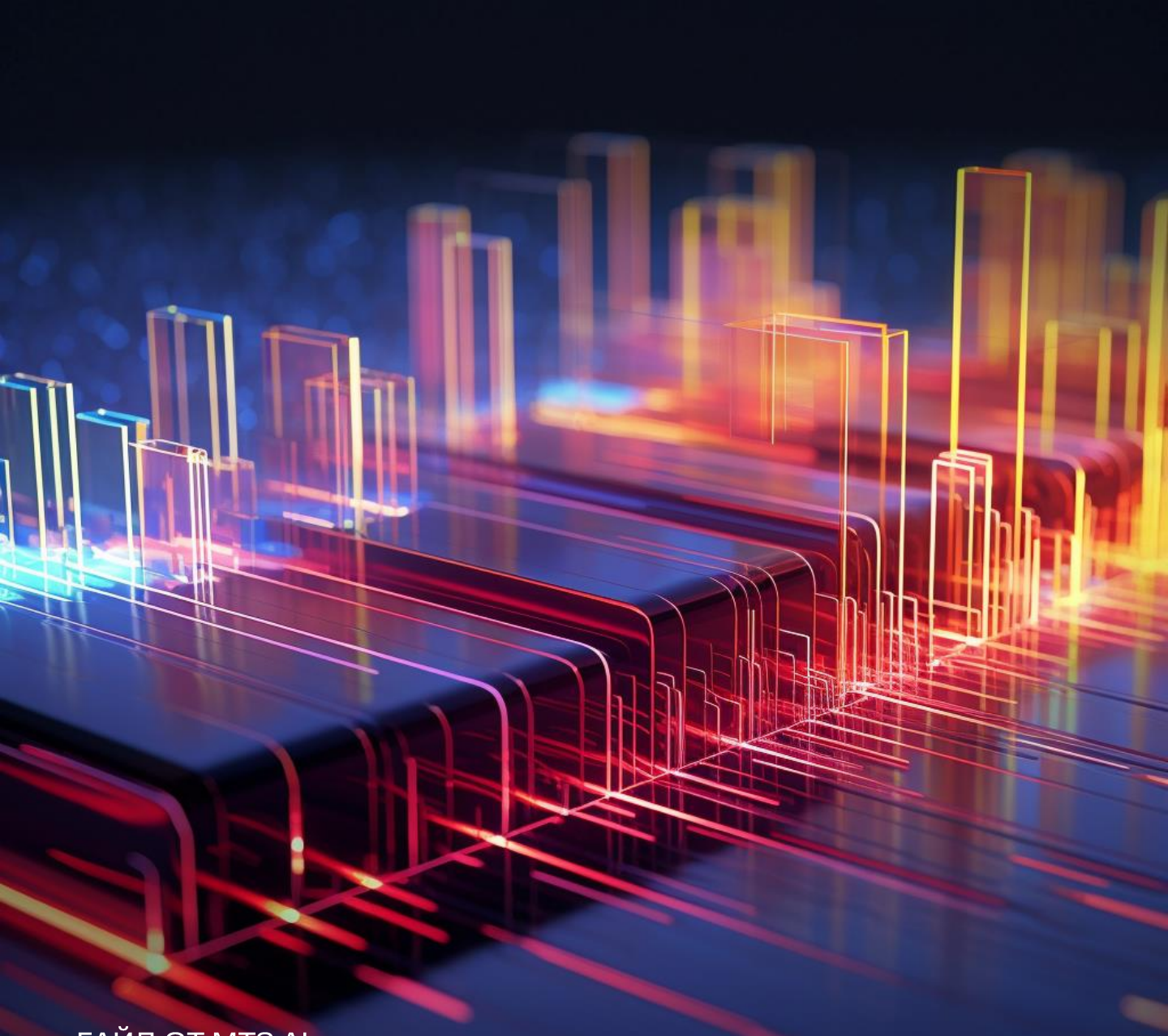

ГАЙД ОТ MTS AI

КАК ВЫБРАТЬ НАИБОЛЕЕ ПОДХОДЯЩИЙ СЕРВИС СИНТЕЗА И РАСПОЗНАВАНИЯ РЕЧИ

Мы подобрали для вас шесть ключевых метрик, на которые стоит ориентироваться бизнесу при выборе сервиса синтеза и распознавания речи.

К каждой из них нужно подходить с умом – не стоит опираться только на один показатель и выбирать наиболее красивые цифры, важно изучать решения в комплексе исходя из ваших бизнес-задач.

На что же стоит обратить внимание прежде всего?

MTS AI

МЕТРИКИ КАЧЕСТВА СИНТЕЗА РЕЧИ

Mean opinion score (MOS)

Для определения качества сервиса вендоры используют усредненную экспертную оценку синтезированной речи (MOS). В рамках исследования большая группа респондентов независимо друг от друга прослушивает фрагменты текста, озвученные нейросетями разных вендоров, и присваивает аудиозаписям оценку по различным критериям: от общего впечатления до правильности произношения слов и употребления пауз. При этом участники исследования не знают, какой образец принадлежит тому или иному вендору.

На основе усредненных показателей сервисам ставят баллы от 1 до 5, где 1 — это совсем неправдоподобное звучание, а 5 — речь, неотличимая от человеческой. У записей речи людей MOS оценивается в 4,5, поэтому значение выше 4 считается высоким. Например, у одного из голосов Audiogram — платформы синтеза и распознавания MTS AI — MOS оценивается в 4.32.

Вендоры проводят такие исследования с определенным интервалом, чтобы следить за развитием модели в динамике.

Каждый разработчик подобных решений сам определяет вопросы для оценки MOS. Они могут отличаться в зависимости от задач использования нейросети. Например, специфические параметры применяются при подсчете MOS для озвучивания IVR: в них особенно важна разборчивость речи.

Стоит понимать, что показатель MOS очень субъективен, и он может отличаться даже в рамках одной AI-модели на разных примерах текста для озвучивания.

Какие показатели обычно исследуются для оценки MOS?

- Общее впечатление
- Естественность звучания речи
- Соответствие интонации тексту
- Корректность расстановки пауз
- Уровень общей зашумленности
- Наличие артефактов (дрожание голоса, растягивание звука и т.д.)
- Высота голоса
- Скорость чтения
- Неправильное произношение слов

Side-by-side (SbS)

Это способ сопоставления двух AI-решений с помощью экспертной оценки. Он используется для сравнения разных версий одного алгоритма синтеза речи или для конкурентного анализа решений различных вендоров. В отличие от MOS, респонденты прослушивают несколько фрагментов одного и того же текста, озвученных разными AI-системами. Однако для оценки образцов речи используются те же параметры: темп голоса, естественность звучания и т.д.

После прослушивания эксперты присваивают синтезированным образцам оценку по шкале от -2 до +2: эти значения показывают, насколько одна запись лучше другой по тому или иному показателю. Как и в случае с MOS, эксперты не знают, какой образец принадлежит тому или иному вендору. По итогам анализа исследователи получают усредненное значение (например, +0,95), по которому они определяют, насколько их решение лучше или хуже альтернативного.

SbS является внутренним инструментом анализа вендоров, однако бизнес тоже может использовать этот способ исследования для выбора наиболее подходящего решения.

Скорость синтеза речи (SPS)

SpS — это метрика скорости, технический показатель, который определяет, какой объем речи будет синтезироваться ежесекундно и какое количество параллельных потоков для этого будет использоваться.

Метрика SpS тесно связана с техническим оборудованием. Например, на одной видеокарте NVIDIA® Tesla®V100 с 16 ядрами CPU и 64 Gb RAM, система способна синтезировать 300 SpS при 30 одновременных подключениях в потоковом режиме.

Соответственно, если заказчику необходима более высокая скорость синтеза речи, то для этого потребуется более мощная техническая инфраструктура. Однако большие значения SpS нужны далеко не всегда, а только при значительных объемах текста для воспроизведения, или, например, если необходимо озвучивать материалы в онлайн-режиме.

МЕТРИКИ КАЧЕСТВА РАСПОЗНАВАНИЯ РЕЧИ

Word Error Rate (WER)

Для сравнения сервисов распознавания речи в основном используются показатель доли неправильно распознанных слов в тексте (WER). Чем он ниже, тем выше точность распознавания.

Использовать WER при выборе вендора стоит очень осторожно, поскольку оценка этого показателя даже у одного и того же разработчика может отличаться в зависимости от качества и содержания датасета, на котором измерялись метрики. AI-модель вендора может быть хуже адаптирована для распознавания речи на ту или иную тему, поэтому при выборе поставщика необходимо анализировать WER нейросети на собственных датасетах. Эксперты MTS AI считают, что достаточно надежными будут оценки, полученные при распознавании датасетов длительностью от 7 часов.

Близким к идеалу считается показатель около 5%, однако в условиях сильной зашумленности и плохого качества исходной записи отличным считается WER до 10%. По оценкам MTS AI, у сервиса распознавания речи Audiogram в незашумленном телефонном канале этот показатель составляет 11-13%, в более шумной обстановке - около 18%.

Во многих случаях крайне низкий (а значит, самый лучший) показатель WER не требуется: например, для аналитики коммуникаций нейросети могут оценить содержание разговора даже с погрешностью в распознавании некоторых слов. Для корректной работы голосовых ботов WER может чуть ниже 25%, так в этом случае основную роль играет не точность распознавания речи, а логика самого голосового ассистента. При выборе вендора с наиболее низким WER предприниматели рискуют приобрести услуги по высокой стоимости при значительном потреблении ресурсов, однако это может почти не сказаться на бизнес-показателях.

Тем не менее добиться близких к нулю показателей WER все же возможно при ограниченном наборе слов, которые потребуется распознать нейросети. Например, когда навигатор принимает команды по конкретной базе адресов, все названия улиц заранее занесены в нейросеть, и их будет просто распознать с высокой точностью.

Lemma Error Rate (LER)

Этот показатель учитывает процент неправильно распознанных лемм – начальных форм слова. В отличие от WER, при оценке LER неправильно распознанное окончание слова не будет считаться ошибкой, поскольку зачастую для работы нейросетей неправильно распознанное окончание некритично. Благодаря этому средний показатель LER, как правило, на 2% ниже WER.

Эта метрика актуальна для распознавания синтетических языков (русский, чешский, немецкий и т.д.), в которых слова образуются с помощью окончаний, суффиксов и т.д. Ее можно использовать для сравнения с оценкой WER на англоязычных датасетах, поскольку в них практически нет окончаний.

Пример:

Если голосовой помощник онлайн-магазина “услышит” фразу “подарок мамы” вместо “подарок маме”, он все равно продолжит разговор по корректной ветке диалогового скрипта. Для таких кейсов можно использовать показатель LER.

Если медицинский работник использует сервис распознавания речи для приема пациентов и диктует показания, ему критически необходимо, чтобы система распознала все слова целиком, чтобы не занести в систему некорректный диагноз. Для таких случаев необходимо ориентироваться на метрики WER.

Semantic Distance

Специально обученная нейросеть может определить смысл предложения в распознанном тексте и оценить, насколько он совпадает по значению с исходным материалом. В результате анализа AI-система формирует оценку смысловой близости распознанного текста к оригиналу.

Это необходимо для оценки критичности ошибок при распознавании речи. Сравните: если при исходной фразе “я очень хочу игрушку” нейросеть распознает текст как “я осень хочу игрушку”, то это не влияет на конечный смысл. Если же ИИ-модель “услышит” фразу как “я не хочу игрушку”, это полностью изменит значение исходного текста.

Пока ИТ-разработчики не создали единую схему подсчета этих метрик, и каждый вендор считает ее по-своему. Чтобы выбрать оптимальное значение Semantic Distance для вашего проекта, в целом нужно понимать, для чего ваша компания будет использовать синтез и распознавание речи, и насколько критичны будут ошибки для этих задач.

Переходите по ссылке, чтобы узнать больше о возможностях Audiogram — платформы синтеза и распознавания речи от MTS AI, протестировать озвучивание текстов и транскрибацию аудио.

Все вопросы по продукту можно задать,
написав письмо по адресу: sales@mts.ai

С уважением,
MTS AI