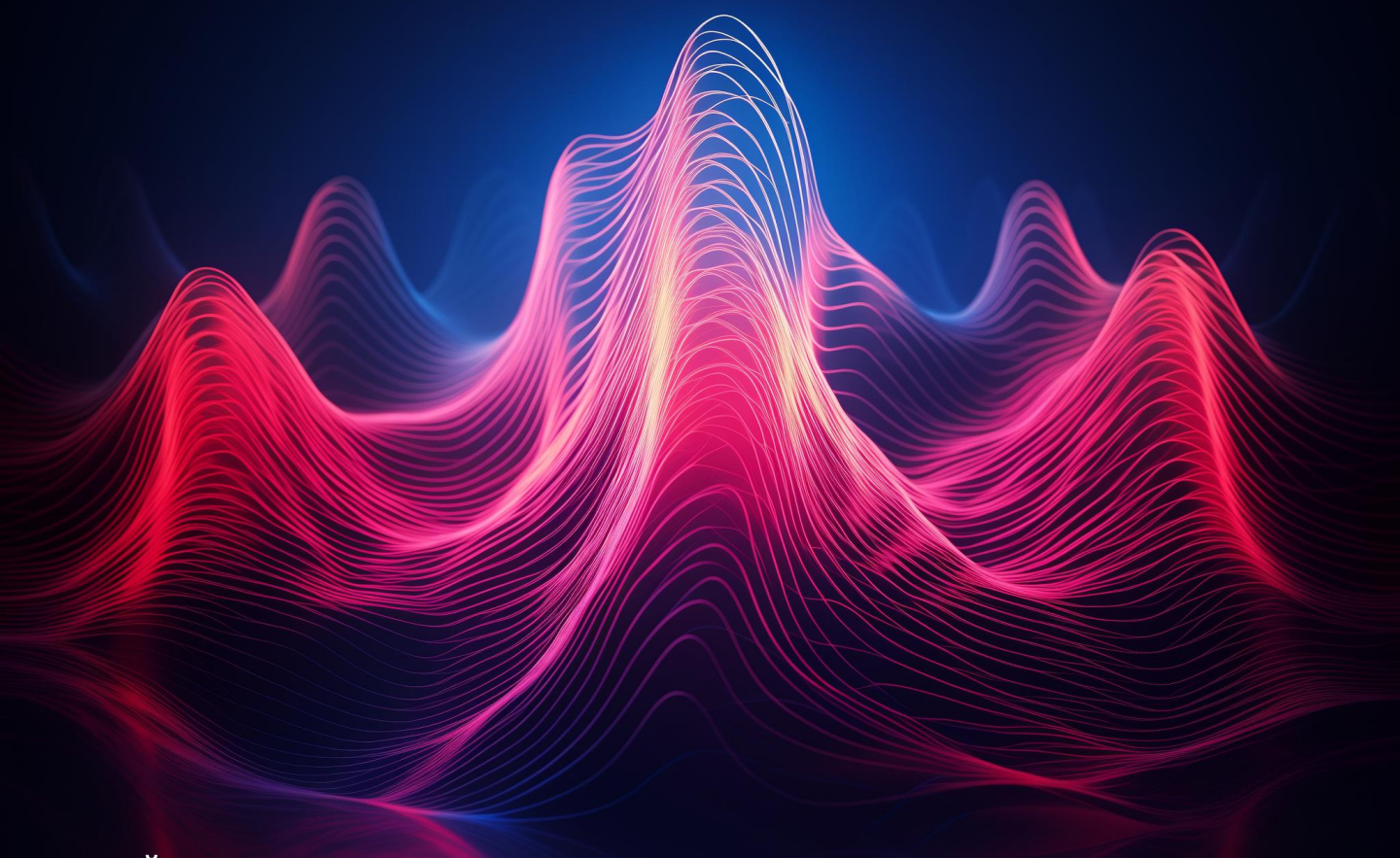


ГАЙД ОТ MTS AI

КАК ОЦЕНИТЬ КАЧЕСТВО ГОЛОСОВОГО БОТА

Компании, сотрудники которых много и регулярно общаются со своими клиентами, стремятся максимально автоматизировать голосовую коммуникацию с помощью ботов. Банки, ритейлеры, медицинские организации, телеком-операторы и другие активно используют их для первичных консультаций клиентов по продуктам и услугам, для приема заказов и технический поддержки, а также для обзвонов – например, для маркетинговых коммуникаций с персональными предложениями.

По экспертным оценкам MIT и Genesys, использование разговорного ИИ позволяет увеличить выручку на 5-10%. Применение ботов ускоряется обслуживание, снижает нагрузку на операторов, освобождает их время для решения более тяжелых задач.

Разработку голосовых помощников под заказ предлагают десятки компаний, однако далеко не все из них универсальны и подходят бизнесу. Специалисты MTS AI подготовили список метрик, на которые стоит ориентироваться при выборе наиболее подходящего решения.

MTS AI

ИЗ ЧЕГО СКЛАДЫВАЕТСЯ ОЦЕНКА БОТА

Выбор решения для голосового бота зависит от конкретных потребностей бизнеса. Сами по себе метрики не показательны: некоторые из них субъективны и считаются по разным методикам, и для каждой задачи нормы одного и того же показателя могут разниться. Поэтому все метрики необходимы анализировать в комплексе и применительно к конкретным целям внедрения ИИ-помощника.

Качество работы бота можно оценить по трем группам критериев: голосовой, коммуникационной и инфраструктурной.

01 Голосовая группа

Влияет качество работы ASR и TTS моделей — автоматического распознавания и синтеза речи.

02 Коммуникационная группа

Коммуникационная часть зависит от возможностей NLP-платформы, с помощью которой создается голосовой бот – формируется логика его работы, прорабатываются сценарии общения с клиентами и так далее.

03 Инфраструктурная группа

Инфраструктурные параметры зависят от вычислительных мощностей и архитектуры систем, которые использует компания: перед запуском бота необходимо проверить совместимость программного обеспечения и железа с минимальными требованиями работы нейросети, или же использовать ее в качестве облачного решения.

MTS AI

КОММУНИКАЦИОННЫЕ МЕТРИКИ БОТА

Для оценки работы бота прежде всего необходимо измерить удовлетворенность клиентов обслуживанием. Соответственно, большинство метрик для коммуникационной платформы основано на отзывах клиентов, а также на анализе их дальнейших действий по итогам разговора (не расторгнут ли они договор и не оставят ли жалобу). Исключительно субъективная оценка будет нерепрезентативна — клиент может быть недоволен недостаточной вежливостью голосового помощника и поставить ему низкую оценку после звонка, хотя бот полностью решил его проблему.

БИЗНЕС-МЕТРИКИ ГОЛОСОВЫХ БОТОВ

При оценке качества голосового бота бизнесу наиболее приоритетен баланс между двумя ключевыми метриками: уровнем удовлетворенности клиента и gross автоматизацией (долей диалогов ботов с клиентами, которые завершились без привлечения оператора). Для поиска этого баланса задействуются все возможности бота: сценарии, модели, запросы к данным о пользователе — в зависимости от них бизнес может калибровать уровень gross автоматизации.

Уровень удовлетворенности клиента: tNPS, CSAT и другие

Для определения уровня удовлетворенности клиентов бизнес может использовать разные традиционные метрики: NPS, tNPS, CSAT, CES и сочетание некоторых из них. Они отличаются не только по методике подсчета, но и по смыслу. Например, при измерении CSAT клиент оценивает работу техподдержки, а при NPS — сам продукт.

Традиционно для оценки бота используется метрика tNPS — она практически идентична NPS, но показывает не общий срез по всем отзывам, а позволяет подробно проверить клиентский опыт каждого человека, прошедшего опрос.

Метрика tNPS вычисляется на основе ответов на вопросы, наподобие «на сколько вероятно, что Вы порекомендуете нашу компанию/продукт другим?». Ответы собираются на шкале от 0 до 10. Оптимальным показателем tNPS читается 6 — он означает, что от применения голосового бота положительного результата больше, чем возможного негатива.

Однако определить подходящий для всех ситуаций показатель удовлетворенности затруднительно: у компаний отличаются не только используемые метрики, но и методика подсчета: выборка респондентов, время для опроса, каналы связи с клиентом. Кроме того, в некоторых компаниях есть заведомо негативные сценарии общения с клиентами, которые отрицательно скажутся на метриках удовлетворенности – однако они не будет говорить о качестве работы бота.

Оценка tNPS зависит от общего уровня обслуживания клиентов. При высоком уровне gross автоматизации на tNPS будет в большей степени влиять работа голосовых ботов, при низком – самих операторов колл-центра. Для более объективного анализа tNPS необходимо сопоставлять со смежными показателями – количеством инцидентов, которые произошли в диалогах с ботами, частотой негативных действий клиентов (например, переход на другого оператора) после разговора с голосовым помощником и так далее.

Gross автоматизация

На текущем уровне развития голосового ИИ многие боты все еще не могут ответить на вопросы клиента так же качественно, как человек. Поэтому статистически чем выше gross автоматизация, то есть чем чаще бот самостоятельно отвечает на вопросы и не переводит собеседника на оператора, тем ниже общая метрика tNPS.

Из-за этого у каждого бизнеса в зависимости от его задач, используемой лексики и вопросов, которые задают операторам, есть своя цена gross автоматизации за 1% tNPS.

Для безусловной удовлетворенности клиентов голосовой помощник должен отвечать исключительно на те вопросы, в которых он полностью уверен в ответе. Однако это приводит к низкому уровню gross автоматизации, и в таком случае бизнес не получит экономии от использования бота. Из-за низкой уверенности в ответе помощник может ошибочно переводить абонентов на операторов по простым вопросам, тем самым почти не снижая нагрузку на колл-центр.

Оптимальным показателем считается 40-50% gross автоматизации: свыше 50% может пострадать клиентский опыт и tNPS, а ниже 40% могут быть потери в бизнес-метриках из-за недостаточной экономии на работе колл-центра.

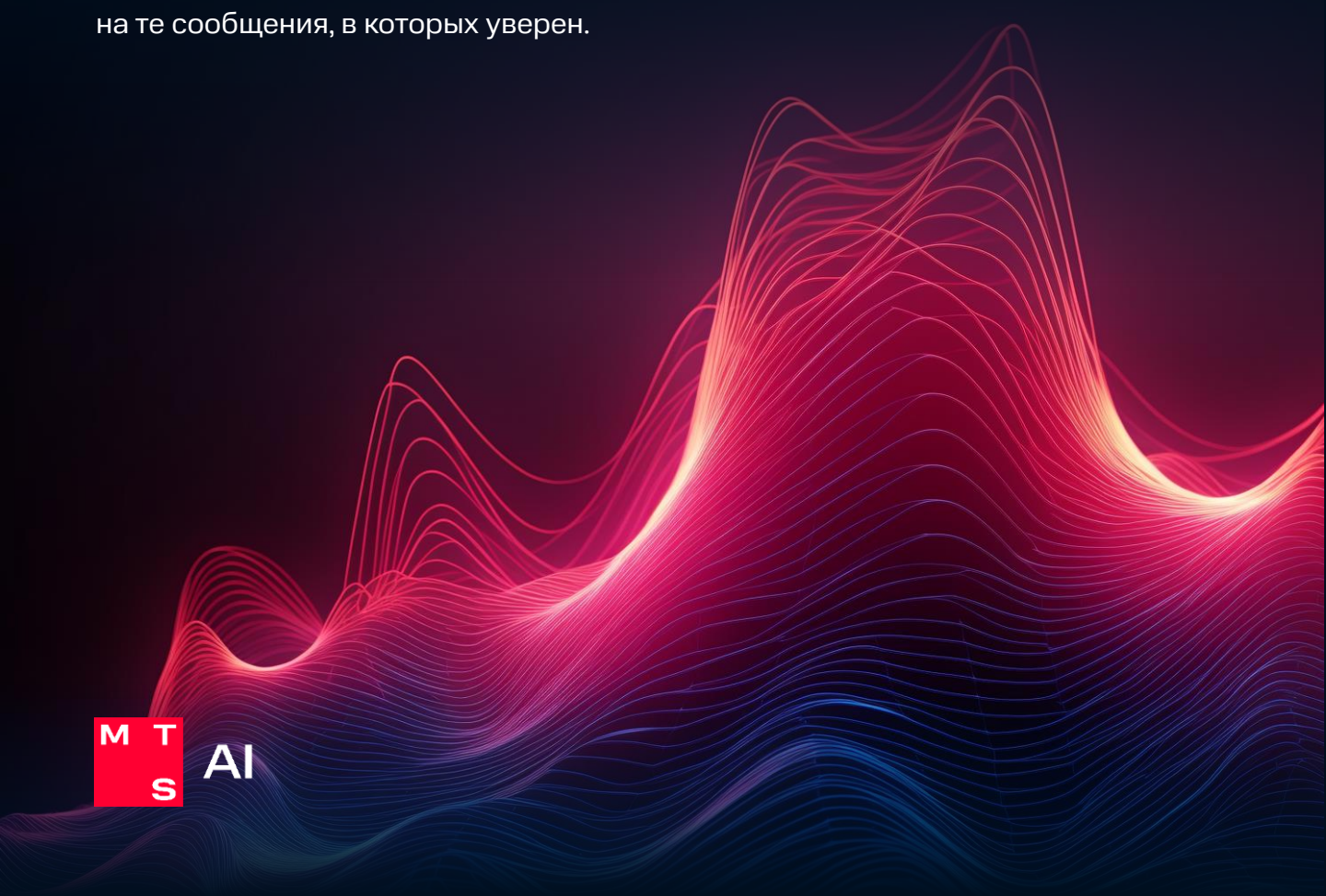
Тем не менее компания может позволить себе более высокий уровень gross автоматизации при корректно обученных алгоритмах нейросетей и правильно сформулированных диалоговых скриптах. Для более конкретного анализа бизнес может использовать показатель gross автоматизации по определенным темам – например, сразу перенаправляя клиента на оператора по потенциально негативным сценариям разговора.

Доля распознанных фраз

Для определения сбалансированных параметров настройки бота используется еще одна метрика – доля корректно распознанных фраз в диалоге. На ней сказывается уровень gross автоматизации и частота самостоятельных ответов бота на вопросы клиента.

> Золотым стандартом доли распознанных фраз считается 85% – при его достижении можно работать над повышением gross автоматизации при сохранении этого показателя.

Для балансировки показателя можно использовать пороги распознавания. Нейросеть, “поняв” запрос клиента, может определить тематику обращения по той или иной категории с уверенностью от 0 до 1. Чем выше пороговое значение “уверенности” модели, тем чаще бот пропускает вопрос и переводит абонента на оператора, тем самым потенциально избегая снижения tNPS. Бизнес может повысить пороговое значение, если ему необходимо увеличить долю корректно распознанных фраз в диалоге, чтобы бот отвечал только на те сообщения, в которых уверен.



ПРОКСИ-МЕТРИКИ ДЛЯ РАЗРАБОТЧИКОВ

Помимо бизнес-метрик, ML-специалисты также могут следить еще за несколькими техническими параметрами. Они напрямую не влияют на показатели компании, но их анализ поможет при калибровке и дополнительном обучении нейросети.

First call retention (FCR)

Эта метрика традиционно используется для оценки работы колл-центров и способности операторов решить вопросы собеседника с первого звонка. Ее же используют для анализа качества бота — чтобы проверить, как часто клиенты возвращаются в колл-центр после общения с голосовым помощником.

Для этого аналитики выделяют диалоги, которые были завершены без участия оператора, и проверяют, сколько клиентов, которые общались с голосовым помощником в рамках этих диалогов, перезванивает в течение определенного интервала — например, от 3 минут до 7 часов, этот период считается ключевым для оценки повторных обращений. Соответственно, чем меньше пользователи обращаются вновь, тем лучше ИИ справляется со своими функциями.

Доля доспрашиваний

Идеально настроенный и обученный бот с первого раза дает клиенту ответ, который удовлетворяет пользователя и содержит достаточное для него количество информации. Считается, что чем реже у него возникают дополнительные вопросы к голосовому помощнику, тем лучше работает нейросеть. При высоком уровне доспрашиваний бизнесу может быть необходимо скорректировать сценарии общения ботов с клиентами — это может положительно повлиять на tNPS.

Во время тестирования нейросети при изменении сценариев можно отследить, в каких ситуациях происходит резкий всплеск количества доспрашиваний. Такие точки указывают на вопросы, на которые бот не знает ответа — например, о новых товарах или услугах. В таком случае бизнесу стоит создать новую ветку диалога и сформировать новый сценарий для бота.

ГОЛОСОВЫЕ МЕТРИКИ БОТА

Метрики синтеза и распознавания речи трудно определить объективно, поскольку многие из них основаны на оценке от группы респондентов. Тем не менее при четком понимании собственных целей они помогут узнать, насколько та или иная нейросеть подходит для интеграции в бизнес.

МЕТРИКИ РАСПОЗНАВАНИЯ РЕЧИ

WER и LER

Для оценки распознавания речи используются две ключевых метрики – WER (Word error rate) и LER (Lemma error rate). Они показывают долю некорректно распознанных слов (WER) или начальных форм слова без учета окончаний (LER). Чем меньше ошибок, тем ниже эти метрики – и тем лучше бот понимает клиентов. Например, показатель 10% WER означает, что из текста в 100 слов 10 из них были распознаны неверно.

Оценка LER используется для компаний, в которых точное распознавание слова по буквам некритично. Например, при подборе тарифа бот поймет, о чем говорит клиент оператора, и без правильных окончаний в речи собеседника. В таком случае можно использовать метрику LER.

Практически идеальным уровнем распознавания считается показатель WER около 5%, однако при сильном шуме или плохом качестве звука хорошим результатом будет считаться и около 10%. Средний показатель LER всегда примерно на 2% ниже WER, потому что неправильно распознанные окончания не будут считаться ошибкой.

При оценке качества распознавания речи не стоит ориентироваться на наиболее низкие метрики – для корректной работы войс-ботов достаточно метрики WER чуть ниже 25% при условии хорошо обученных алгоритмах работы нейросети.

M T S AI

Semantic Distance

ML-специалисты работают над еще одной новой метрикой семантической близости – Semantic Distance. С помощью нейросети аналитики могут оценить смысловую близость различных аспектов текста к оригиналу. Таким образом она способна определить критичность тех или иных ошибок в решениях для распознавания речи.

Например, если клиент скажет “я очень хочу тариф”, а нейросеть распознает этот текст как “я осень хочу тариф”, смысл принципиально не изменится, и такая ошибка не будет критичной. Однако ошибочная расшифровка “я не хочу тариф” может принципиально изменить дальнейшую логику диалога.

МЕТРИКИ СИНТЕЗА РЕЧИ

MOS

Ключевой метрикой для определения качества синтеза речи считается Mean opinion score (MOS) – усредненная экспертная оценка синтеза. Для ее определения группа респондентов анализирует несколько аудиозаписей, которые синтезированы нейросетями от разных вендоров.

Каждый член группы оценивает записи по шкале от 1 до 5 по различным характеристикам: по общему впечатлению, правильности произношения слов и другим параметрам.

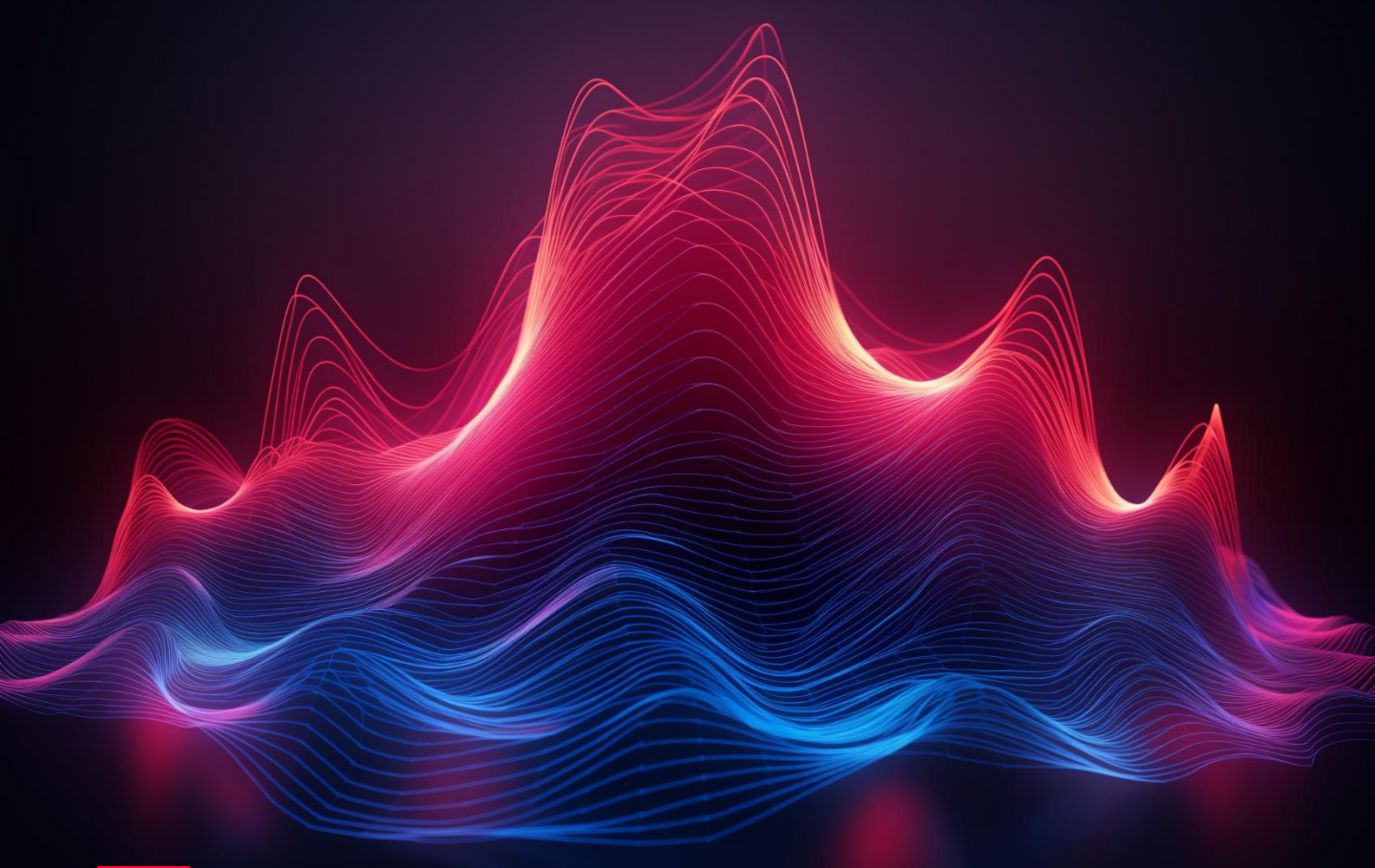
Усредненная оценка по всем параметрам и будет показателем MOS для каждой конкретной нейросети. При этом создать идеальную модель практически невозможно – даже у настоящей человеческой речи MOS обычно не превышает 4,5 баллов.

SbS

Для сопоставления качества работы нейросети относительно других ML-моделей также используется метрика Side-by-Side. Она также формируется на базе мнений респондентов, однако в этом случае они выставляют свои оценке по шкале от -2 до +2 в зависимости от того, насколько одно решение превосходит другое по тем или иным параметрам.

SPS

Компаниям с небольшими вычислительными ресурсами также важно учитывать технический показатель скорости синтеза речи (SPS). Высокие показатели этой метрики необходимы для быстрой обработки большого объема аудиоматериала: например, в контакт-центре или при общении с виртуальным ассистентом. Однако в большинстве случаев в высокой скорости нет необходимости – она почти не скажется на бизнес-метриках, но будет расходовать ресурсы инфраструктуры и обойдется значительно дороже.
Для примера, 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.



M T S AI