**Оценка качества синтеза речи модел****ью**

**Parler-TTS-mini-jenny-30H**

Климов О.Д.

# **Конкретизация цели и задач**

В рамках работы рассматривается вопрос оценивания качественных характеристик синтеза речи text-to-speech моделью Parler-TTS-mini, дообученной на датасете Jenny. Был выделен следующий набор характеристик:

1. Корректность и разборчивость синтезируемой речи;
2. Согласованность синтезируемого голоса с голосом диктора;
3. Корректность интерпретации текстовых промптов (“description” – описание, определяющее уровень шума, эха и пр.) моделью;
4. Качество звучания синтезируемых аудиодорожек;
5. Натуральность синтезируемой речи;
6. Робастность модели.

Метрики, используемые для оценки характеристик, описаны в главах 1-5.

# **Введение**

Parler-TTS-mini-jenny-30H **[1]** – text-to-speech модель, полученная в результате дообучения модели Parler-TTS-mini v0.1 **[2]** на датасете Jenny **[3]**, содержащем 30 часов аудиозаписей с голосом ирландского диктора Дженни и транскрипции к ним. Записи обладают высоким качеством и практически лишены посторонних шумов. С помощью набора утилит Data-Speech **[4]** датасет был дополнен текстовыми описаниями речи, подаваемыми на вход модели в качестве дополнительных промптов.

Поскольку в датасете Jenny нет изначального разбиения на обучающую и тестовую выборки, для оценивания модели были взяты 100 примеров, записанных одним и тем же диктором, из датасета LibriTTS-R **[5]**, характеризующихся следующими признаками:

1. **Сходство между голосами Дженни и выбранного диктора:** критично для метрик, характеризующих натуральность сгенерированной речи и согласованность сгенерированного голоса с голосом диктора;
2. **Разборчивость и корректность речи:** сравнительная оценка корректности и разборчивости синтезированной речи предполагает, что оригинальные записи, рассматриваемые как Ground Truth, обладают этими свойствами в полной мере;
3. **Большая дисперсия качества, длительности и содержания аудиозаписей**: необходима для оценки робастности модели.

Отобранные примеры были дополнены текстовыми описаниями, сгенерированными с помощью Data-Speech. Для оценок восприимчивости модели к промптам и качества звучания синтезированной речи будут использованы взятые из датасета и дополненные вручную написанными описаниями транскрипции записей.

# **Оценка корректности и разборчивости синтезированной речи**

## Word error rate

Разборчивость слов и корректный синтез их звучания можно оценить по правильности их последующего распознавания. Word error rate (WER) при этом вычисляется как отношение кол-ва ошибок в распознанном тексте (вставленных, удаленных или замененных слов) к кол-ву слов в оригинальной транскрипции. В работах **[5]** и **[6]** для распознавания были использованы ASR-модели (в **[5]** оценивание производилось с использованием двух ASR-моделей с разным механизмом действия). В связи с тем, что часть оригинальных и синтезированных записей может содержать шумы или эхо, в настоящей работе в качестве метрик, помимо, непосредственно, WER-syn сгенерированной записи, были взяты корреляции между WER-syn и значениями уровней шума и эха, а также разность (не абсолютная) между WER-syn и WER-o, где положительные значения говорят о том, что синтезированная речь более разборчивая, чем оригинальная, отрицательные – об обратном, а околонулевые значения – о равенстве оценок, независимо от уровня шума, эха или ошибок ASR-модели.

# **Оценка согласованности синтезируемого голоса с голосом диктора**

## SIM

Speaker similarity reflects the in-context learning capability of zero-shot TTS models. We utilize WavLM-TDNN5 (Chen et al., 2022) to extract speaker embedding vectors from the reference speech prompt and the generated speech. The cosine distance between them is then calculated to measure speaker similarity, with the range of [-1, 1]. We compute SIM-r and SIM-o, where SIM-r measures the similarity between the synthesized speech and the speech prompt reconstructed from the mel-spectrogram with the vocoder, whereas SIM-o measures the similarity with respect to the original speech prompt. Considering that SIM-r is not comparable among systems using different acoustic tokens, we recommend referring to SIM-o.

<https://github.com/microsoft/UniSpeech/tree/main/downstreams/speaker_verification>

Coherence: This is measured by the similarity between the embedding of generated speech and that of the audio context, where different embedding models would reflect coherence of different attributes. VALL-E proposed to use WavLM-TDCNN speaker embedding model [Chen et al., 2022], which maps an audio clip to a fixed dimensional vector, to measure voice similarity. We consider the same model to compare with VALL-E. In particular, VALL-E reports similarity with respect to resynthesized audio context by its vocoder (Encodec-decoder), which we call SIM-resyn (SIM-r). SIM-resyn is not comparable across models using different vocoders. Hence, we advocate for computing similarity against the original audio context, which we call SIM-orig (SIM-o).

## MCD

Mel cepstral distortion - a measure of how different two sequences of mel cepstra are (generated samples and ground truth samples). It is used in assessing the quality of parametric speech synthesis systems, including statistical parametric speech synthesis systems, the idea being that the smaller the MCD between synthesized and natural mel cepstral sequences, the closer the synthetic speech is to reproducing natural speech (quality of synthetic speech)

<https://github.com/SamuelBroughton/Mel-Cepstral-Distortion/blob/master/mel-cepstral-distortion.ipynb>

# **Оценка восприимчивости модели к текстовым промптам**

Для тестовой выборки были взяты транскрипции из исходного датасета, дополненные случайными значениями pitch\_mean, pitch\_std, snr, c50, speaking\_rate из диапазонов [100.0, 300.0], [3.0, 80.0], [10.0, 80.0], [10.0, 80.0], […] соответственно, по которым с помощью Data-Speech были сгенерированы текстовые описания. В качестве метрик были рассмотрены корреляции между оригинальными pitch\_mean, pitch\_std, snr, c50, speaking\_rate и значениями, вычисленными с помощью Data-Speech для сгенерированных записей.

# **Оценка качества звучания**

Our final objective evaluation aims to quantify the audio fidelity of our model when asked to produce audio with “excellent recording quality” or similar terms. Here, we use the recently proposed Torchaudio Speech Quality and Intelligibility Measures [34]. This model provides a reference-less estimate of Wideband Perceptual Estimation of Speech Quality (PESQ), Short-Time Objective Intelligibility (STOI), and Scale-Invariant Signal-to-Distortion Ratio (SI-SDR). Using 20 test sentences and descriptions from LibriTTS-R, we run these metrics on outputs from our model, Audiobox (using the public website interface9), and the ground truth audio

<https://arxiv.org/pdf/2304.01448>

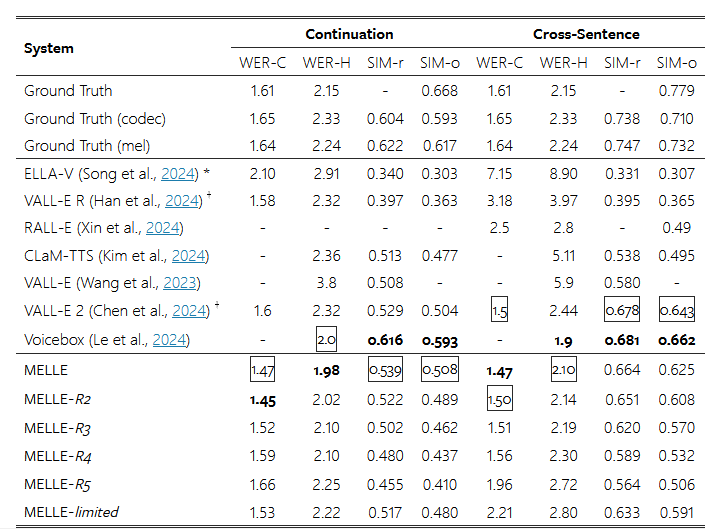
# **Оценка натуральности синтезируемой речи**

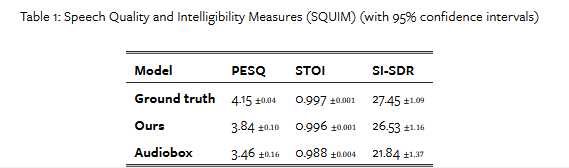
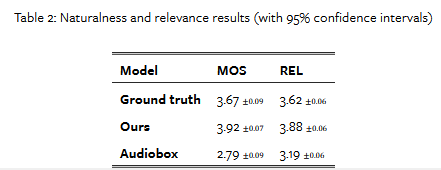
UTMOS - a surrogate objective metric of MOS (speech quality)

<https://github.com/sarulab-speech/UTMOSv2>

# **Выводы**

[Здесь будет размещена сводная таблица с метриками, а также будут описаны выводы]

Результаты, полученные в работах […] при оценке других TTS-моделей:



# **Ссылки на ресурсы**

1. <https://huggingface.co/parler-tts/parler-tts-mini-jenny-30H>
2. <https://huggingface.co/pharaouk/parler_tts_mini_v0.1>
3. <https://huggingface.co/datasets/reach-vb/jenny_tts_dataset>
4. <https://github.com/huggingface/dataspeech?ysclid=m1o7jpwvh1819142243>