arXiv:2402.08093v2 [cs.LG] 15 Feb 2024

BASE TTS: Lessons from building a billion-parameter Text-to-Speech model on 100K hours of data

Mateusz Łajszczak* **Guillermo Cámbara*** **Yang Li*** **Fatih Beyhan***

Arent van Korlaar* **Fan Yang** **Arnaud Joly** **Álvaro Martín-Cortinas†**

Ammar Abbas **Adam Michalski** **Alexis Moinet** **Sri Karlapati** **Ewa Muszyńska**

Haohan Guo **Bartosz Putrycz** **Soledad López Gambino** **Kayeon Yoo**

Elena Sokolova **Thomas Drugman**

Amazon AGI ‡

Abstract

We introduce a text-to-speech (TTS) model called BASE TTS, which stands for **B**ig **A**daptive **S**treamable TTS with **E**mergent abilities. BASE TTS is the largest TTS model to-date, trained on 100K hours of public domain speech data, achieving a new state-of-the-art in speech naturalness. It deploys a 1-billion-parameter autoregressive Transformer that converts raw texts into discrete codes ("speechcodes") followed by a convolution-based decoder which converts these speechcodes into waveforms in an incremental, streamable manner. Further, our speechcodes are built using a novel speech tokenization technique that features speaker ID disentanglement and compression with byte-pair encoding. Echoing the widely-reported "emergent abilities" of large language models when trained on increasing volume of data, we show that BASE TTS variants built with 10K+ hours and 500M+ parameters begin to demonstrate natural prosody on textually complex sentences. We design and share a specialized dataset to measure these emergent abilities for text-to-speech. We showcase state-of-the-art naturalness of BASE TTS by evaluating against baselines that include publicly available large-scale text-to-speech systems: YourTTS, Bark and TortoiseTTS. Audio samples generated by the model can be heard at `https://amazon-ltts-paper.com/`.

1 Introduction

Generative deep learning models are progressing at a rapid pace. Natural Language Processing (NLP) and Computer Vision (CV) are undergoing a fundamental shift from specialized models with supervised training, to generalized models that can achieve miscellaneous tasks with limited explicit instruction [1]. In NLP, tasks such as question answering, sentiment analysis and text summarization can now be performed by a large language model (LLM) which was not specifically targeted for these

*Main contributors. Study conception and design: Mateusz, Guillermo; data collection: Mateusz, Fatih; experiments: Mateusz, Arent, Fatih, Guillermo; evaluation and analysis: Mateusz, Yang, Arent, Guillermo; draft manuscript preparation: Mateusz, Yang, Arent, Guillermo; supervision: Mateusz,Yang, Arent.

†Universidad Politécnica de Madrid. Work performed while an intern at Amazon.

‡Correspondence: amazon-ltts-paper@amazon.com

tasks [2–6]. In CV, pre-trained models that learn from hundreds of millions of image-caption pairs have achieved top performance on image-to-text benchmarks [7–9], while delivering remarkably photo-realistic results in text-to-image tasks [10–14]. This progress has been enabled by Transformer-based architectures [15] that drive improvements using many orders of magnitude more data than previous models. Similar advances are now occurring in Speech Processing and Text-to-Speech (TTS), with models leveraging thousands of hours of data that push synthesis ever closer towards human-like speech. Some of these models, described in depth in Section 5, rely on causal language modeling tasks, like AudioLM [16] or VALL-E [17], whereas others use non-causal modules, such as SoundStorm [18] or SpeechX [19], or diffusion decoders [20, 21].

Until 2022, leading Neural TTS models were almost exclusively trained on a few hundreds of hours of recorded audio [22–26]. Such systems can create well-enunciated speech that is occasionally expressive for the target speakers, but typically cannot generalize beyond the small amount of training data to render ambiguous and complex texts with truly expressive spoken performance [27–29]. To achieve such higher levels of expressiveness, TTS systems historically had to rely on labeled datasets for specific speech phenomena; and even so, achieving human-like prosody for certain types of textual inputs has remained elusive [30]. For example, in English, compound nouns and questions are notoriously hard to render correctly without accurate syntactic parsing and semantic understanding [31].

In this paper, we introduce BASE TTS: Big Adaptive Streamable TTS with Emergent abilities. It is a multi-lingual and multi-speaker Large TTS (LTTS) system trained on around 100K (doubling previous high in [17]) hours of public domain speech data. BASE TTS follows the approach of casting TTS as a next-token-prediction problem [16, 17, 21], inspired by the success of LLMs. This approach is usually applied in combination with large amount of training data to achieve strong multi-lingual and multi-speaker capabilities (e.g. one-shot voice cloning). Our goal is to improve general TTS quality and study how scaling affects the model's ability to produce appropriate prosody and expression for challenging text inputs, similar to how LLMs acquire new abilities through data and parameter scaling, a phenomenon known as "emergence" or "emergent abilities" in the LLM literature [32, 33]. [32] defines *emergent abilities of large language models* as "abilities that are not present in smaller-scale models but are present in large-scale models;" for example, they show that on a range of few-shot tasks, model capability stays at a low level from 10^{18} to 10^{22} training FLOPs, but makes a drastic jump from 10^{22} to 10^{24}. To test the hypothesis that this also holds for LTTS, we propose an evaluation scheme to assess potential emergent abilities in TTS, identifying seven categories that are challenging from the literature [27–31]: compound nouns, emotions, foreign words, paralinguistics, punctuations, questions, and syntactic complexities. See more details in Section 3.3.

We design BASE TTS to model a joint distribution of text tokens followed by discrete speech representations, which we refer to as speechcodes. Discretization of speech through audio codecs [34–36] is central to our design, as it enables the direct application of methods developed for LLMs, which underlie recent works on LTTS [16–21, 37]. Specifically, we model speechcodes using a decoder-only autoregressive Transformer with a cross-entropy training objective. Despite its simplicity, this objective can capture complex probability distributions of expressive speech, thus alleviating the oversmoothing problem seen in early neural TTS systems [38]. As an implicit language model, BASE TTS is also observed to make a qualitative jump in prosody rendering once a large enough variant is trained on sufficient data.

Further, we propose speaker-disentangled speechcodes that are built on top of a WavLM [39] Self-Supervised Learning (SSL) speech model. We follow [16] which introduces semantic tokens constructed by discretizing activations of an SSL model. We extend this approach to better control information captured by the speechcodes. Our strategy is to limit the responsibilities of the autoregressive speechcode predictor ("speechGPT") to segmental contents, prosody, and duration, while designating a separate, speechcode-to-waveform decoder (called "speechcode decoder") with the reconstruction of speaker identity and recording conditions. We show that this convolution-based speechcode decoder is compute-efficient and reduces the whole-system synthesis time by over 70% compared to the baseline diffusion-based decoder.

Our main contributions are summarized as follows:

I. We introduce BASE TTS, which to our knowledge is the largest TTS model to date, featuring 1B parameters and trained on a dataset consisting of 100K hours of public domain speech

data. In subjective evaluations, BASE TTS performs better than publicly available LTTS baseline models.

II. We demonstrate how scaling BASE TTS to larger dataset and model sizes improves its capability to render appropriate prosody for complex texts. To this end, we develop and make available an "emergent abilities" testset that can serve as a subjective evaluation benchmark for text understanding and rendering of large-scale TTS models. We report performance on different variants of BASE TTS over this benchmark, showing monotonic improvement in quality with increased dataset size and parameter count.

III. We introduce novel discrete speech representations that are built on top of a WavLM SSL model, intended to capture only phonemic and prosodic information of the speech signal. We demonstrate that these representations outperform the baseline quantization method. We also show that they can be decoded to high quality waveforms with a simple, fast, and streamable decoder, despite high level of compression (only 400 bits/s).

2 BASE TTS

2.1 Overview

Similar to recent works in speech modeling, we adopt an LLM-based method for the TTS task (Figure 1). We consider a dataset $\mathcal{D} = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=0}^{N}$, where $\mathbf{y}$ is an audio sample and $\mathbf{x} = \{x_1, \cdots, x_T\}$ is the corresponding text transcription. The audio $\mathbf{y} = \{y_1, \cdots, y_S\}$ is represented by a sequence of S discrete tokens (speechcodes), learnt using a separately trained speech tokenizer. We use a Transformer-based autoregressive model with parameters ϕ in order to learn the joint probability of the text and audio sequences:

$$p(\mathbf{y}, \mathbf{x}) = p(\mathbf{y}|\mathbf{x})p(\mathbf{x}) = \prod_{s=1}^{S} p(\mathbf{y}_s|\mathbf{y}_{<s}, \mathbf{x}; \phi) \prod_{t=1}^{T} p(\mathbf{x}_t|\mathbf{x}_{<t}; \phi). \quad (1)$$

The predicted speech tokens are concatenated with speaker embeddings and decoded into waveforms using a separately trained speechcode decoder consisting of linear and convolutional layers.

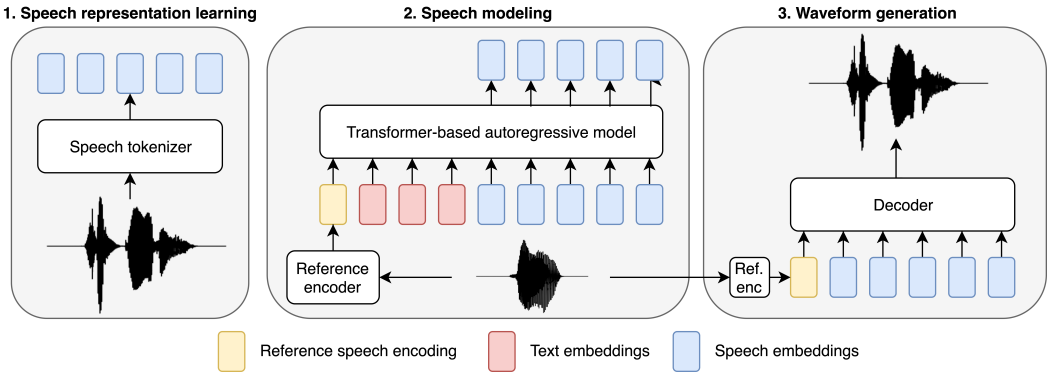

Figure 1: An overview of **BASE TTS**. The speech tokenizer (1) learns a discrete representation, which is modeled by an autoregressive model (2) conditioned on text and reference speech. The speechcode decoder (3) converts predicted speech representations into a waveform.

2.2 Discrete speech representations

Discrete representation is foundational to the success of LLM, but identifying a compact and informative representation is less obvious in speech than in text, and less explored. For BASE TTS, we first experiment with a Vector Quantized Variational Autoencoder (VQ-VAE) baseline [34], where an autoencoder based architecture reconstructs mel-spectrograms through a discrete bottleneck, as described in Section 2.2.1. VQ-VAE has been a successful paradigm in speech and image representation, and especially as a unit of modeling for TTS [21, 40].

In Section 2.2.2, however, we introduce a novel method of learning speech representations through WavLM-based speechcodes. In this method, we discretize features extracted from a WavLM SSL model to reconstruct a mel-spectrogram. We apply additional loss functions to promote speaker disentanglement, and compress the resulting speechcodes using Byte-Pair Encoding (BPE) [41] to reduce sequence length, enabling us to model longer audio with Transformers.

Both representations are compressed (325 bits/s and 400 bits/s respectively) to allow more efficient autoregressive modeling compared to popular audio codecs (e.g. 6k bits/s in [17]). With this level of compression, we aim to remove information from speechcodes that can be reconstructed during decoding (speaker, audio noise, etc.) to ensure that the capacity in speechcodes is primarily dedicated to encoding phonetic and prosodic information.

2.2.1 Autoencoder-based speech tokens

Our baseline discretization method is a VQ-VAE trained to reconstruct mel-spectrograms. The encoder and decoder are convolutional neural networks with residual connections, which downsample the speech representation to a frequency of 25Hz. To (partially) disentangle speaker information from the speech representations, we introduce a global reference encoder [42]. This encoder learns a fixed-size utterance-level representation, which is concatenated to the speechcodes before reconstructing with the VQ-VAE decoder.

From informal listening, we find that speechcodes produced by the autoencoder-based speech tokenizer still contain speaker information. This motivates us to develop representations with improved speaker disentanglement.

2.2.2 WavLM-based speechcodes

We aim to develop speechcodes that contain phonetic and prosodic information, but which are disentangled from speaker identity, recording conditions, and other spurious features in the audio signal. To this end, we introduce a speech tokenizer based on features extracted from a pretrained WavLM model [39], further trained with losses that encourage disentangling the speaker identity. Our approach similar to the one introduced in [43] with modifications that reduce bitrate of the codes. The overall architecture of the speech tokenizer is shown in Figure 2.

We first pass the waveform through the WavLM model and extract the hidden states. These hidden states are then passed through separate content and speaker linear regressors. The output of these regressors is then fed into a convolutional residual encoder [44]. The content encodings are passed through a vector quantization module that outputs one speechcode per one WavLM frame (i.e. 20ms of speech).

The speaker encodings are passed through a Transformer-based speaker extractor [15] to obtain the speaker embeddings. The model only extracts, and we only use non-specific features that cannot be used for identification.

The speaker embeddings are concatenated with the speechcodes, and decoded into a spectrogram using a convolutional decoder. We then compute L1 distance between decoded and target spectrograms and use it as the reconstruction loss. While L1 is not the optimal reconstruction objective, we prioritize representations that are conducive for autoregressive modeling [45], and demonstrate accordingly that the final audio quality can be kept high when this learned representation is decoded with our speechcode decoder, in Section 2.4. The speaker embeddings are used in a contrastive loss, maximizing the similarity between samples from the same speaker and minimizing it for those from different speakers [46]. Furthermore, we maximize the cosine distance between the speaker embeddings and embeddings obtained by passing the output of the content regressor through the frozen speaker extractor and applying gradient reversal [47]. We hypothesize that this encourages disentanglement between content and speaker information.

In addition to better disentanglement of speaker information, we also believe that using features from a pretrained WavLM model as input (as opposed to a jointly learnt audio encoding) keeps speechcodes more robust to recording conditions. Our intuition is that WavLM was trained with data augmentation to encourage disentanglement from background noise. The total loss is given by a weighted combination of these losses, in addition to the commitment loss for the vector quantizer:

$$L = L_{\text{recon}} + \alpha L_{\text{commitment}} + \beta L_{\text{contrastive}} + \gamma L_{\text{cosine}} \tag{2}$$

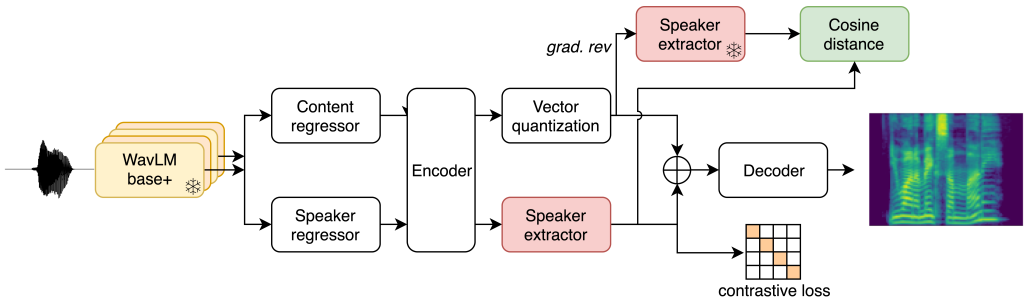

Figure 2: WavLM-based speech tokenizer. The proposed architecture encourages disentanglement of speaker and content information.

2.2.3 Byte-pair encoding on speechcodes

The WavLM-based speech representations are learned at a frequency of 50 Hz to ensure that they offer enough resolution in time to be able to discriminate between single phonemes even in fast-paced speech (the average length of English phonemes tends not to dip below 20ms [48]). We apply Byte-Pair Encoding [41] to reduce the average sequence length of speechcodes by around 40%, in order to mitigate the quadratic memory increases in Transformers sequence length and minimize complexity for the autoregressive model. Similar to the BPE of texts, we iteratively join the most frequent pairs of speechcodes into new tokens and add them to the vocabulary until a pre-determined vocabulary size is reached.

2.3 Autoregressive speech modeling (SpeechGPT)

We train a GPT2-architecture autoregressive model [49] that we call "SpeechGPT" to predict the speechcodes conditioned on text and reference speech. The reference speech conditioning consists of a randomly selected utterance from the same speaker, which is encoded to a fixed-size embedding. The reference speech embedding, text, and speechcodes are concatenated into a single sequence that is modeled by a Transformer-based autoregressive model. We use separate positional embeddings and separate prediction heads for text and speech.

We train the autoregressive model from scratch, without pretraining on text (e.g. as done in [50]). In order to retain textual information to guide prosody, we also train SpeechGPT with an objective to predict the next token in the text portion of the input sequence, so that speechGPT is partially a text-only LM. We apply a lower weight to this text loss compared to the speech loss.

2.4 Waveform generation

Our baseline uses a diffusion-based spectrogram decoder and a separately trained UnivNet [51] vocoder, as proposed in [21]. Diffusion-based TTS decoding can generate high-quality speech, but it suffers from slow inference and cannot generate samples incrementally - This lack of streamabilility forces us to obtain the audio output for the entire sequence in one go. Furthermore, the diffusion model in [21] predicts spectrograms, and requires a separately trained vocoder to generate audio, complicating the training and inference pipeline.

Our proposed decoder, inspired by [23], is trained in an end-to-end fashion to predict waveforms. Our variant uses speechcodes as an input to the model instead of phoneme encodings and prosody latents. Additionally, to make the model more scalable, we replace LSTM layers [52] with convolutional ones to decode an intermediate representation. The output of a HiFi-GAN based decoder block [53] is fed into BigVGAN vocoder [54] to predict the waveform. In training, we use the same set of adversarial and non-adversarial losses as [23]. We call our proposed system a "speechcode decoder" and depict it in Figure 3. In addition to simplifying the overall system, we hypothesize that training the decoder and vocoder end-to-end yields higher-quality speech.

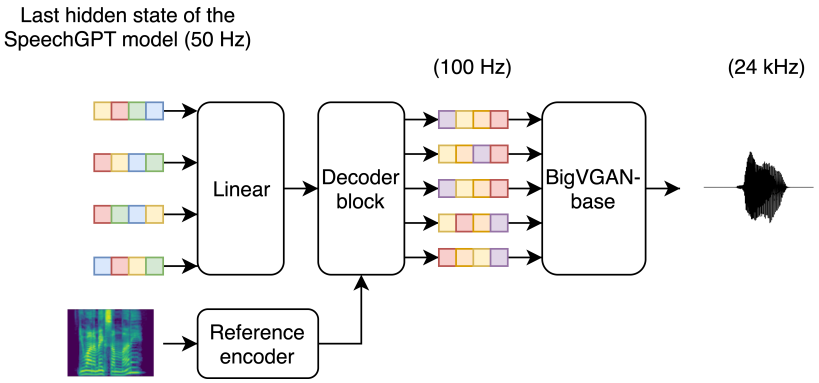

Figure 3: Speechcode decoder. The last intermediate representations of the SpeechGPT model are upsampled by a factor 2 in a decoder block. BigVGAN-base vocodes these representations into the waveform.

In practice, instead of speechcodes, the speechcode decoder takes as an input the last hidden state of the autoregressive Transformer. We do so because the dense latent representation provides much richer information than a single speechcode, following [21]. During training, we feed the text and target codes to the trained SpeechGPT (with parameters frozen) and then condition the decoder on the last hidden state. Feeding the last hidden states of SpeechGPT helps improving the segmental and acoustic quality of speech, but also couples the decoder to a specific version of SpeechGPT. This complicates experimentation because it forces the two components to be always built sequentially. This limitation needs to be addressed in future work.

3 Experimental setup

We design and conduct experiments to validate the architecture of BASE TTS and its quality, capabilities and compute performance. First, we compare model quality achieved by autoencoder-based and WavLM-based speechcodes. We then evaluate the two methods of acoustically decoding the speechcodes: a diffusion-based decoder and a speechcode decoder. Having completed these architectural ablations, we assess the emergent abilities of BASE TTS in 3 variants of dataset sizes and model parameters, assessed by an expert linguist. Further, we run subjective MUSHRA tests to measure naturalness, as well as automated intelligibility and speaker similarity measurements. We also report the quality of speech comparison against other open-source text-to-speech models.

3.1 Dataset

In order to test our hypothesis that abilities emerge with the scale of data, we set out to train with the largest speech dataset for our largest LTTS model. We create a dataset consisting of 100K hours of unlabeled, public domain speech data. The overall dataset is dominated by English data (over 90%), followed by German, Dutch, Spanish. We first download mp3 files from the Web, and resample them into 24 kHz mono LPCM files, with 16 bits per sample encoded as signed-integers. The vast majority of the dataset is recorded in non-studio conditions and contain noises, and we avoid additional signal processing or denoising in order to test our model's ability to generate clean speech from noisy background data.

Next, we use an ASR model combined with Voice Activity Detection (VAD) to perform speech segmentation and Automatic Speech Recognition. The ASR model splits the entire speech into 30-second or shorter fragments. We then perform VAD to split sentences longer than 20 seconds, and reassemble them into longer sentences during training as long as their total length does not exceed 40 seconds. We thus expose BASE TTS to speech segments which are between 0 and 40 seconds long, containing one or more sentences, so that the model becomes robust to both very short inputs, and learns from longer context [55].

To generate the transcript for the dataset, we initially relied on the ASR transcription. We found that ASR tends to generate a smaller set of punctuations than their natural occurrences. For example, it

over-generates commas instead of colons and semicolons, and rarely generates parenthesis. We then perform partial text restoration (similar to [56]) by: 1) searching for source texts from the Internet and tying them to each recording; 2) matching the ASR transcript with the source text sentence by sentence, replacing the former with latter if their difference is small enough (length no longer/shorter by a factor of 3 and within a certain edit distance). We restore around 1/2 of total texts to the 'original', resulting in an 300% increase in quotation marks of all kinds and 20000% increase in parentheses in the transcript. This led to our models no longer having frequent acoustic artefacts on parentheses and other 'rare' punctuations.

3.2 Training and hyperparameters

We train the models in three steps. Due to the use of discrete speech representations, we rely on the cross-entropy validation loss on $\sim$50 hours held-out from training data to guide our training, restarting, and stopping decisions. All models are trained on clusters of AWS P3dn instances with NVIDIA® V100 GPUs.

First, we train two speech tokenizer variants: a VQ-VAE tokenizer as in Section 2.2.1 and a WavLM tokenizer as in Section 2.2.2. Due to the overwhelming presence of English data, we train both variants on a subset of the data in Section 3.1 by keeping all data from non-English languages, but capping the amount of speech from individual English speakers at 200 hours - excess data from the speaker are randomly discarded. We use a codebook size of 256 for the WavLM speechcodes and 8196 for the VQ-VAE ones. The WavLM speechcodes are subsequently compressed with the BPE algorithm and a vocabulary of size 8192. We first optimize the reconstruction loss on ground truth spectrograms, sanity-checking the reconstructed audio on "minority" languages such as Polish (0.2% of total data) to ensure sufficient speechcode coverage.

Second, we train SpeechGPT on the entire training dataset with random initial weights. The inputs to the model are: random reference speech segment from a speaker, text and corresponding target speech segments. We convert the reference speech segment to a mel-spectrogram and feed it to the reference encoder. The target speech segment is tokenized by either VQ-VAE or WavLM, forming the target speechcodes. We concatenate the encoded reference with input text embeddings and speechcodes embedding, and feed them to the autoregressive Transformer. The model is optimized with cross-entropy loss on both text tokens (weight 0.01) and speechcodes (weight 1.0).

Our autoregressive Transformer structure is largely identical to the GPT-2 language model [49]. For most of the experiments reported below ("BASE-large"), we train a 32-layer decoder-only transformer with masked self-attention heads (1536-dimension states and 24 attention heads). For the feed-forward networks, we use 6144-dimension inner states. We use Adam [57] optimization with a max learning rate of 3.0e-4 and weight decay of 0.03. The learning rate was increased linearly from zero over the first 10000 updates and annealed to 1.5e-4 using a cosine schedule.

To understand the importance of data and model size, we build two other versions (BASE-small & BASE-medium) of SpeechGPT with WavLM speechcodes as described in Section 2.2.2, with increasing number of parameters and speech data in Table 1.

Table 1: Ablation experiments on speechGPT with 3 data sizes and parameters

Attributes	**BASE-small**	**BASE-medium**	**BASE-large**
Data amount	1K hours	10K hours	100K hours
Parameters in SpeechGPT	150 million	400 million	980 million
Attention heads	12	16	24
Transformer layers & dims	16 layers, 768 dims	30 layers, 1024 dims	32 layers, 1536 dims
Feed-forward dims	3072	4096	6144
Speechcode	wavLM	wavLM	wavLM
Decoder	Speechcode decoder	Speechcode decoder	Speechcode decoder

Finally, we train the speechcodes decoder described in Section 2.4 with a modified log-likelihood-maximization GAN training scheme borrowed from [23]. Total parameter size for this module is around 150 million.

3.3 Evaluation methodology

Three types of tests are considered to assess the model quality:

I. **MUSHRA (Multiple Stimuli with Hidden Reference and Anchor)** among competing systems by native listeners in US English and Spanish. We use a MUSHRA scale ranging from 0 to 100. We do not place an Upper/Lower anchor or ask listeners to rate those at 0 or 100. This way, we are able to get more ratings on target systems on a fixed evaluation budget. For each MUSHRA test reported, 25-50 testers provide ratings on 50-100 text snippets that range between 5-40 seconds long, such that we achieve about 17-20 hours of total effective listening time per system per test.

II. **Linguistic expert evaluation of "emergent abilities".** To gauge the ability of BASE TTS to achieve finer understanding of the text, we hand-created an "emergent abilities testset" in English with 7 categories of texts: Questions, Emotions, Compound Nouns, Syntactic Complexities, Foreign Words, Paralinguistics, and Punctuations. In Table 2, we present an example from each category, and how a linguistic expert rates the TTS output on a discrete 3-point scale. These sentences are designed to contain challenging tasks - parsing garden-path sentences [58], placing phrasal stress on long-winded compound nouns [59], producing emotional or whispered speech, or producing the correct phonemes for foreign words like "qi" or punctuations like "@" - none of which BASE TTS is explicitly trained to perform. Our hypothesis is that as BASE TTS increases in model capacity and trains over more data, the model will start to acquire these abilities, following evidence that scaling in these dimensions begets qualitative ability jumps [1, 32, 33]. We share the full testset in Appendix A.

III. **Automated objective evaluations** to test TTS robustness, especially issues with missed, duplicated or hallucinated synthesis [20], and speaker similarity. We use an ASR model to compute the Word Error Rate (WER) by comparing the testset text (ground truth) against the ASR output from the synthetic speech. In addition, we employ a speaker verification model [1] fine-tuned on WavLM features [39] to obtain speaker embeddings from the original recordings and synthetic speech, and then we compute the cosine distance between them to get a speaker similarity metric (SIM).

For subjective evaluations, we report average MUSHRA scores with 95% confidence intervals. Furthermore, to determine the significance of differences between two systems, we conduct a t-test; if the p-value is < 0.05, we consider the difference significant. To aid visualization we mark statistically significantly better systems with **bold**.

4 Results

4.1 VQ-VAE speechcode vs. WavLM speechcodes

We conduct MUSHRA evaluations on 6 US English and 4 Spanish (Castilian and US) speakers in order to test comprehensively the quality and generalisability of the two speech tokenization approaches, using both held-out test data for speakers seen in the training ("seen" condition) and unseen speakers ("one-shot" condition). In terms of average MUSHRA scores for English voices, VQ-VAE and WavLM based system are on par (VQ-VAE: 74.8 vs WavLM: 74.7, the difference is statistically non-significant). However, for Spanish voices, WavLM based model outperforms the VQ-VAE one (VQ-VAE: 73.3 vs WavLM: 74.7) in a statistically significant way. Note that English data comprises around 90% of our dataset, while Spanish data only 2%. We hypothesize that speechcode improvements are more critical for low-resource languages while sheer data volume can make up for imperfect representations. Further verification of this hypothesis needs to be addressed in future work. Since WavLM-based system performs at least as well or better as the VQ-VAE baseline, we use it to represent BASE TTS in further experiments. In Table 3, we show the results categorized by speaker to provide additional details on per-speaker performance.

[1] `https://huggingface.co/microsoft/wavlm-base-plus-sv`

Table 2: Emergent abilities testset by category and evaluation criteria.

Categories	Example sentence	Evaluation criteria
Compound Nouns	The Beckhams decided to rent a charming stone-built quaint countryside holiday cottage.	1 = fails to recognise compound nouns 2 = fails to realise the phrasal stress naturally 3 = natural phrasal stress
Emotions	"Oh my gosh! Are we really going to the Maldives? That's unbelievable!" Jennie squealed, bouncing on her toes with uncontained glee.	1 = no audible emotions 2 = emotion present but insufficient 3 = correct emotion recognition and appropriate rendering
Foreign Words	Mr. Henry, renowned for his mise en place, orchestrated a seven-course meal, each dish a pièce de résistance.	1 = pronounces foreign words with incorrect anglicized pronunciation 2 = applies foreign accent but not entirely correctly 3 = correct rendering in the intended language or accepted anglicized reading
Paralinguistics	"Shh, Lucy, shhh, we mustn't wake your baby brother," Tom whispered, as they tiptoed past the nursery.	1 = no recognition of paralinguistic keywords such as "shhh" or "phew" 2 = clear intention to render keywords distinctly, but rendering unnatural 3 = natural rendering, e.g. making speech voiceless on "shhh" and other whispered speech

Punctuations	She received an odd text from her brother: 'Emergency @ home; call ASAP! Mom & Dad are worried...#familymatters.'	1 = glitches on uncommon punctuations such as # or &
		2 = no glitch but incorrect rendering
		3 = no glitch and correct pausing and verbalization, e.g. @ as "at".

Questions	But the Brexit question remains: After all the trials and tribulations, will the ministers find the answers in time?	1 = intonation pattern incorrect 2 = intonation pattern largely correct but with minor flaws 3 = correct intonation
Syntactic Complexities	The movie that De Moya who was recently awarded the lifetime achievement award starred in 2022 was a box-office hit, despite the mixed reviews.	1 = failure to parse the syntax correctly 2 = parses the syntax largely correctly but the rendering is not entirely natural 3 = parsing correct and rendering natural

Table 3: Results of MUSHRA evaluation comparing VQ-VAE and WavLM speechcodes. We report mean scores for 10 US English & Spanish speakers and highlight statistically significant winners with **bold**.

Speaker	Language	One-shot or seen	VQ-VAE	WavLM
Male speaker A	US English	seen	76.3 ± 0.6	76.3 ± 0.6
Male speaker B	US English	one-shot	76.3 ± 0.6	76.3 ± 0.6
Male speaker C	US English	one-shot	73.6 ± 0.6	73.4 ± 0.6
Female speaker A	US English	seen	74.5 ± 0.7	74.5 ± 0.7
Female speaker B	US English	seen	73.8 ± 0.7	74.0 ± 0.7
Female speaker C	US English	one-shot	74.1 ± 0.6	73.9 ± 0.6
Male speaker D	Castilian Spanish	seen	72.5 ± 0.9	**74.4** ± 0.9
Male speaker E	US Spanish	one-shot	73.9 ± 0.8	74.4 ± 0.8
Female speaker D	Castilian Spanish	seen	77.0 ± 0.7	77.0 ± 0.7
Female speaker E	Castilian Spanish	one-shot	68.5 ± 1.0	**72.4** ± 1.0

4.2 Diffusion-based decoder vs. speechcode decoder

BASE TTS simplifies over the baseline diffusion-based decoder by proposing an end-to-end speechcode decoder, as described in Section 2.4. Our approach offers streamability and 3X improvement in inference speed. In order to ensure that this approach does not degrade quality, we run an evaluation of the proposed speechcode decoder against the baseline. Table 4 presents the results of MUSHRA evaluation we conducted for 4 US English and 2 Spanish speakers. For 4 voices out of 6, the BASE TTS variant with speechcode decoder outperforms the diffusion-based baseline in terms of average MUSHRA score. For the remaining speakers the difference is not statistically significant. We conclude that the speechcode decoder is the preferred approach, as it does not degrade quality and for most of the voices it brings quality improvements, while offering faster inference. Our results suggest that combining two powerful generative models for speech modeling is redundant and can be simplified by dropping the diffusion decoder.

Table 4: Results of MUSHRA evaluation comparing diffusion-based decoder and speechcode decoder. We report mean scores for 6 US English & Spanish speakers and highlight statistically significant winners with **bold**.

Speaker	Language	One-shot or seen	Diffusion-based decoder	Speechcode decoder
Male speaker A	US English	seen	73.3 ± 0.8	**74.9** ± 0.8
Male speaker B	US English	one-shot	74.6 ± 0.7	**75.5** ± 0.7
Female speaker A	US English	seen	77.9 ± 0.7	77.4 ± 0.7
Female speaker C	US English	one-shot	69.2 ± 0.8	**71.1** ± 0.8
Male speaker D	US Spanish	seen	77.3 ± 0.8	77.0 ± 0.8
Female speaker E	Castilian Spanish	one-shot	72.8 ± 1.0	**74.1** ± 1.0

4.3 Emergent abilities - Data and model size ablation

In this section, we report on our verification of the hypothesis that data and parameter scaling in LTTS brings qualitatively different results, analogous to training LLMs from 10^{22} to 10^{24} tokens, when LLMs "suddenly" start to master few-digit addition, recognise words in context at above chance level, and transcribe speech with expert phonetic alphabet.

We perform both MUSHRA and Linguistic expert judgement of "Emergent abilities," as described in 3.3, on 2 American English voices. We report all scores by BASE-small, BASE-medium and BASE-large systems, as outlined in Table 1. In the MUSHRA results in Table 5, we observe that speech naturalness improves significantly from BASE-small to BASE-medium, but less so from BASE-medium to BASE-large - the difference is statistically significant only in Male speaker A.

We report the results of the linguistic expert judgement for the three systems, with their average score for each category, in Figure 4. First, we see a universal jump from BASE-small to BASE-medium across categories. BASE-small appears fundamentally unable to interpret emotions, paralinguistics, and foreign words (average score < 1.25, score lower bounded at 1), and never reaches an average score of 1.75 in any category. By contrast, at BASE-medium, the model has mastered compound nouns and makes a significant jump in all categories; BASE-medium never performs below an average score of 1.75 in any category. From BASE-medium to BASE-large, we observe continued but diminishing improvement in all categories except compound nouns, where the performance has saturated. Combined with the findings from naturalness MUSHRA, we believe that scaling GPT-based TTS model from 1000+ to 10000+ hours and model size from 100 million to 500 million is the point at which "emergent abilities" [32] start to occur for our TTS.

Examining results category by category, we observe that Emotions and Paralinguistics remain the most challenging tasks for all model variants, and even BASE-large performs at around an average score of 2.0, which indicates that model can recognise and react to relevant keywords, but the prosody quality remains imperfect to expert judgement. By contrast, BASE-medium can perform close to ceiling on English Compound Nouns, while trailing BASE-large slightly in most of the other categories (Foreign Words, Punctuations, Questions, Syntactic Complexity). We remain hopeful that further scaling and injection of textual knowledge from text-only LLM can help us close remaining performance gaps.

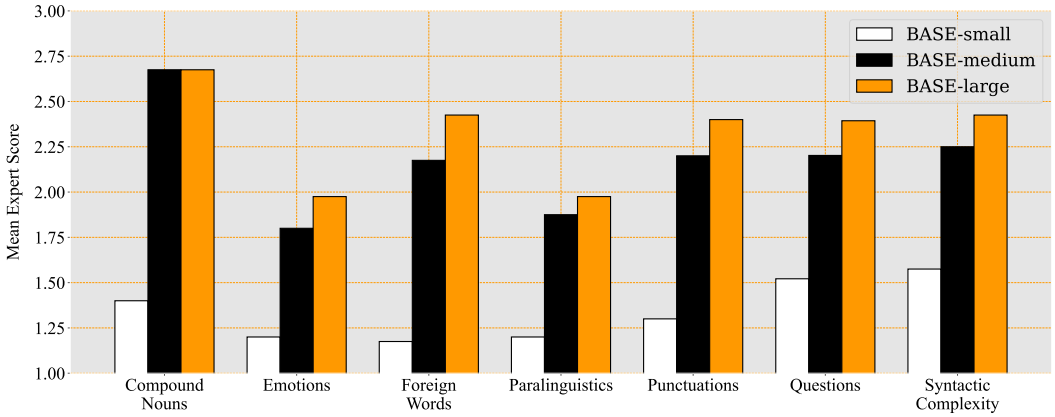

Figure 4: Linguist expert evaluations per system: BASE TTS - small/medium/large. Results are presented for the seven proposed tasks, by computing the mean of the expert scores over 20 sentences in each category.

Table 5: Results of MUSHRA evaluation comparing 3 data-parameter settings on emergent abilities testset. We report mean scores for 2 US English speakers with statistically significant winners highlighted in **bold**.

Speaker	BASE-small	BASE-medium	BASE-large
Male speaker A	61.5 ± 0.7	70.5 ± 0.6	**72.2** ± 0.5
Female speaker A	68.1 ± 0.6	**72.0** ± 0.6	**72.3** ± 0.6

4.4 BASE TTS vs. industry baselines

We select three industry baselines with publicly available pre-trained models: YourTTS [60], Bark[2], and Tortoise [21]. We conduct the comparison exclusively in one-shot setting, using 10-second reference clip for two US English speakers which BASE TTS was not previously trained or evaluated on.

As seen in Table 6, BASE TTS obtains significantly higher naturalness scores in MUSHRA than baseline systems, among two US English speakers tested.

[2] `https://github.com/suno-ai/bark`

We further conduct objective evaluation on WER and speaker similarity on 10 speakers and 307 samples for each, reported in Table 7. Overall, BASE TTS produces the most natural speech, the least amount of misalignment with input text, and the most similar speech to the reference speaker, followed relatively closely by Tortoise on naturalness and WER. Bark and YourTTS perform far worse on naturalness and WER, although Bark is relatively close to BASE TTS in speaker similarity.

Table 6: Results of MUSHRA evaluations comparing BASE TTS with industry baselines. We report mean scores for 2 US English speakers with statistically significant winners highlighted in **bold**.

Speaker	BASE TTS	Tortoise	Bark	YourTTS
Male speaker A	**71.7** $\pm$ 1.0		46.0 $\pm$ 1.0	47.6 $\pm$ 1.0
Male speaker A	**73.7** $\pm$ 0.9	68.8 $\pm$ 0.8		
Female speaker A	**67.8** $\pm$ 1.0		49.7 $\pm$ 1.0	39.7 $\pm$ 1.0
Female speaker A	**68.2** $\pm$ 1.0	58.5 $\pm$ 1.0		

Table 7: Word Error Rate (WER) and Speaker Similarity (SIM) metrics in percentages (%) for our models and industry baselines. We report means and highlight winners with **bold**.

	BASE TTS	Bark	Tortoise	YourTTS
WER $\downarrow$	**6.5**	19.2	8.0	16.3
SIM $\uparrow$	**92.7**	91.7	90.3	90.1

4.5 Synthesis efficiency gain due to speechcode decoder

The speechcode decoder is capable of streaming, i.e. generating speech incrementally. Combining this feature with the autoregressive SpeechGPT gives our system a first-byte latency as low as 100ms - only a few decoded speechcodes are sufficient to produce intelligible speech. This minimal latency contrasts with the diffusion-based decoder that requires the entire speech sequence (one or multiple sentences) to be generated in one go, where first-byte latency equals total generation time. Further, we observe that the speechcode decoder makes the overall system 3X times more compute efficient compared to the diffusion baseline. We run a benchmark where we generate 1000 utterances of around 20 seconds in duration on a NVIDIA® V100 GPU with batch size of 1. On average, it takes 69.1 seconds for 1-billion-parameter SpeechGPT with the diffusion decoder to complete synthesis, but only 17.8s for the same SpeechGPT with the speechcode decoder.

5 Related work

Text-to-speech as language modeling. Casting TTS problem as next token prediction has gained popularity in recent years due to how easy it is to scale language models to large data and model sizes.

The first model using this approach was TortoiseTTS [21] released in 2022. It combines a GPT2-style speechcode language model with a diffusion decoder and an off-the-shelf vocoder, achieving remarkable few-shot capability. VALL-E [17] followed a similar approach: the model is scaled to 60k hours of speech data and uses a pre-trained audio encoder EnCodec for speechcode extraction [36]. VALL-EX [61] replicated VALL-E with 70k hours, with an additional cross-lingual speech-to-speech translation task in English and Chinese by using target language token as prompt. VioLa [62] extended VALL-EX with both text-to-text translation and speech-to-text recognition tasks. SpeechX [19] extended VALL-E by adding noise suppression, target speaker extraction, clean and noisy speech editing, and speech removal tasks.

While VALL-EX, VioLA and SpeechX are versatile models reporting improved performance on word error rate and speaker similarity, no effects on core TTS aspects such as naturalness or prosody improvement were reported. VALL-E reported significant improvements over industry baselines, but we are unable to compare as the model is not publicly available. Here, we propose a multi-lingual and multi-speaker TTS leveraging 100K hours of data, with strong proven naturalness results in English

and Spanish. Evaluations show improved speech naturalness compared to Tortoise-TTS. Qualitative linguistic analysis shows "emergent abilities" [32] - BASE TTS can render complex prosody patterns, taking cues from texts without explicit labels for emotions.

Discrete speech representations. In audio generative AI, the first acoustic encoders used in GPT-TTS are VQ-VAE [34], SoundStream [35] and EnCodec [36], which are deep learning audio compression techniques producing discrete acoustic token based on vector quantization. These acoustic encoders are shown to be superior to Mel-spectrograms [63, 64], but ignore semantic information and are unnecessarily complex due to lack of disentanglement [65]. AudioLM [16] combines SoundStream tokens with semantic BERT tokens [66]. SpeechTokenizer [67] and RepCodec [65] aim to capture semantics by adding self-supervised embedding prediction-based losses [68]. BASE TTS builds acoustic tokens by leveraging semantic information from self-supervised embeddings and disentangling speaker information from the acoustic tokens. Then, it applies byte-pair encoding on these tokens to reduce memory requirement, allowing the model to train on longer sequences.

LTTS simplifies modeling. Since Tacotron 2 [22], a successful TTS paradigm has been to separate speech creation into three sub-systems: (1) a Frontend responsible for text normalization, grapheme-to-phoneme conversion and SSML tag handling e.g. to mark emphasis [69], (2) a Mel-spectrogram generator, inferring the amplitudes of the speech signal and responsible for the model expressivity, and (3) a dedicated Vocoder [70, 71], generating the waveform by inferring the phase information. Expert knowledge can be induced to improve its performance such as using pre-computed features [25], coarse or fine-grained prosody information ([72], and style control [73]. It has also proven successful with architectures such as Transformers [25, 74], Flows [24], and Diffusion [26].

These systems require a complex pipeline. Any error made during an earlier stage is propagated to the next. BASE TTS and other LTTS models are breaking these limitations. We simplify the data preparation by requiring only a large amount of audio, where texts can be obtained from a speech-to-text system, and require no phoneme extraction. GPT-based architectures [49] have flourished by enabling versatile prompt-based task formulation, integration of expert feedback [75], and use as a foundational model with quick fine-tuning [76, 77] e.g. on a specific task, a set of speakers or a new locale. BASE TTS shows that an end-to-end approach can achieve high expressivity on a few audio examples and in a multilingual setting, marking a high bar of quality in Spanish.

Contextual and emotional TTS. This usually requires separate, dedicated TTS sub-systems. Multiple approaches rely on predicting prosody representations using contextualized word embeddings [78–81]. While prosody predictors of these models usually generate appropriate prosody, they often ignore strong text cues that would force a dramatic change e.g. in emotions or speech cues like shouting or whispering. Context-aware emotional TTS systems [82–86] are even more limited. They usually favor a text-based emotion predictor coupled with an emotion controlable TTS system. These systems require high-quality recordings with a forced speaking style and annotated audio and text data, limiting their usefulness due to the sheer number of emotions expressible in human speech [87]. BASE TTS benefits from being a language model which is both acoustic and semantically/syntactically aware. In this paper, we systemize an approach to produce and evaluate emergent contextual understanding of the text with a wide range of styles, without requiring supervised training or annotation.

BASE TTS has data efficiency built-in. Low-resource TTS has focused on reducing the required amount of high quality training data through e.g. voice conversion [88], data generation [89], modeling techniques [69], or recording script optimizations [90]. This stems from the early difficulty of ingesting a high volume of data from multiple speakers in potentially different styles or languages. A step up in that direction is zero-shot inference such as in Bark [3] and [60, 91], which aim to clone a voice with only a few seconds of recordings. Due to leveraging a substantially larger dataset, a bigger model, and a dedicated speaker encoder, BASE TTS and the recent LTTS models set a new standard in data efficiency.

[3] `https://github.com/suno-ai/bark`

6 Conclusion

We introduced BASE TTS, a GPT-style TTS system using novel SSL-based speechcodes as an intermediate representation and a speechcode decoder that offers an efficient, streamable alternative to diffusion. This is the largest model of its kind known to us, both in terms of parameters and training data. We demonstrated new state-of-the-art TTS results against baselines including Tortoise, Bark and YourTTS. We proposed a new way to measure textual understanding of TTS models, and showed that LTTS models built with 10K hours of data and 400 million parameters start to exhibit an advanced grasp of text that enables contextually appropriate prosody. From BASE TTS's strong performance on English and Spanish, we caught a first glimpse of a multilingual TTS approach that achieves high expressiveness, adapts to textual clues, is data efficient, uses only public domain data, and works for streaming TTS usecases such as voicing LLM outputs. Our approach points towards potential Scaling Laws [92] of LTTS models, where an even larger amount of speech and other (text, image) data are needed to support multimodal objectives [93] and to break new grounds in TTS.

Our approach still contains some limitations: a) BASE TTS occasionally produces hallucinations and cutoffs, where we produce either extra or incomplete audio than intended by the text. This is an inherent problem with the autoregressive LM approach, made worse by the misalignment between audio data and the ASR-generated text; b) Selecting the right discrete representation for GPT-style TTS is crucial. More research is needed to establish how different properties of speechcodes translate into end-to-end system quality. We only report results for one speechcode configuration and leave more comprehensive study for future work.

7 Ethical statements

BASE TTS is a high-fidelity model capable of mimicking speaker characteristics with just a few seconds of reference audio, providing many opportunities to enhance user experiences and support under-resourced languages. An application of this model can be to create synthetic voices of people who have lost the ability to speak due to accidents or illnesses, subject to informed consent and rigorous data privacy reviews. However, due to the potential misuse of this capability, we have decided against open-sourcing this model as a precautionary measure. Further, we acknowledge the impact of speech data composition on the ability of the model to express the speech of linguistic, ethnic, dialectal, and gender minorities [94–96]. We advocate for further research to a) quantify the impact of data composition; b) identify methods to combat potential biases and foster inclusivity in voice products.

References

[1] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of artificial general intelligence: Early experiments with gpt-4, 2023.

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[3] OpenAI. Gpt-4 technical report, 2023.

[4] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

[5] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[6] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[7] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[8] Ajinkya Tejankar, Maziar Sanjabi, Bichen Wu, Saining Xie, Madian Khabsa, Hamed Pirsiavash, and Hamed Firooz. A fistful of words: Learning transferable visual models from bag-of-words supervision. *arXiv preprint arXiv:2112.13884*, 2021.

[9] Sheng Shen, Chunyuan Li, Xiaowei Hu, Yujia Xie, Jianwei Yang, Pengchuan Zhang, Zhe Gan, Lijuan Wang, Lu Yuan, Ce Liu, et al. K-lite: Learning transferable visual models with external knowledge. *Advances in Neural Information Processing Systems*, 35:15558–15573, 2022.

[10] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[11] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[12] Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. In *International Conference on Machine Learning*, pages 16784–16804. PMLR, 2022.

[13] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

[14] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022.

[15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *CoRR*, abs/1706.03762, 2017. URL `http://arxiv.org/abs/1706.03762`.

[16] Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matt Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, et al. Audiolm: a language modeling approach to audio generation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2023.

[17] Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. Neural codec language models are zero-shot text to speech synthesizers. *arXiv preprint arXiv:2301.02111*, 2023.

[18] Zalán Borsos et al. SoundStorm: Efficient parallel audio generation. *arXiv preprint arXiv:2305.09636*, 2023.

[19] Xiaofei Wang, Manthan Thakker, Zhuo Chen, Naoyuki Kanda, Sefik Emre Eskimez, Sanyuan Chen, Min Tang, Shujie Liu, Jinyu Li, and Takuya Yoshioka. Speechx: Neural codec language model as a versatile speech transformer. *arXiv preprint arXiv:2308.06873*, 2023.

[20] Kai Shen, Zeqian Ju, Xu Tan, Yanqing Liu, Yichong Leng, Lei He, Tao Qin, Sheng Zhao, and Jiang Bian. Naturalspeech 2: Latent diffusion models are natural and zero-shot speech and singing synthesizers. *CoRR*, abs/2304.09116, 2023. doi: 10.48550/arXiv.2304.09116. URL `https://doi.org/10.48550/arXiv.2304.09116`.

[21] James Betker. Better speech synthesis through scaling. *arXiv preprint arXiv:2305.07243*, 2023.

[22] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, R. J. Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. Natural TTS synthesis by conditioning wavenet on mel spectrogram predictions. *CoRR*, abs/1712.05884, 2017. URL http://arxiv.org/abs/1712.05884.

[23] Syed Ammar Abbas, Sri Karlapati, Bastian Schnell, Penny Karanasou, Marcel Granero Moya, Amith Nagaraj, Ayman Boustati, Nicole Peinelt, Alexis Moinet, and Thomas Drugman. ecat: An end-to-end model for multi-speaker tts & many-to-many fine-grained prosody transfer. In *Interspeech 2023*, 2023. URL https://www.amazon.science/publications/ecat-an-end-to-end-model-for-multi-speaker-tts-many-to-many-fine-grained-prosody-transfer.

[24] Jaehyeon Kim, Jungil Kong, and Juhee Son. Conditional variational autoencoder with adversarial learning for end-to-end text-to-speech. In *International Conference on Machine Learning*, pages 5530–5540. PMLR, 2021.

[25] Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. Fastspeech 2: Fast and high-quality end-to-end text to speech. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL https://openreview.net/forum?id=piLPYqxtWuA.

[26] Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. Grad-tts: A diffusion probabilistic model for text-to-speech. In *International Conference on Machine Learning*, pages 8599–8608. PMLR, 2021.

[27] Andreas Triantafyllopoulos, Björn W Schuller, Gökçe İymen, Metin Sezgin, Xiangheng He, Zijiang Yang, Panagiotis Tzirakis, Shuo Liu, Silvan Mertes, Elisabeth André, et al. An overview of affective speech synthesis and conversion in the deep learning era. *Proceedings of the IEEE*, 2023.

[28] Marie Tahon, Gwénolé Lecorvé, and Damien Lolive. Can we generate emotional pronunciations for expressive speech synthesis? *IEEE Transactions on Affective Computing*, 11(4):684–695, 2018.

[29] Bastian Schnell. Controllability and interpretability in affective speech synthesis. Technical report, EPFL, 2022.

[30] Xu Tan, Tao Qin, Frank Soong, and Tie-Yan Liu. A survey on neural speech synthesis. *arXiv preprint arXiv:2106.15561*, 2021.

[31] Tom Kenter, Manish Sharma, and Rob Clark. Improving the prosody of rnn-based english text-to-speech synthesis by incorporating a bert model. In *INTERSPEECH 2020*, pages 4412–4416, 2020.

[32] Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.

[33] Taylor Webb, Keith J Holyoak, and Hongjing Lu. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, pages 1–16, 2023.

[34] Aaron van den Oord, Oriol Vinyals, and koray kavukcuoglu. Neural discrete representation learning. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/7a98af17e63a0ac09ce2e96d03992fbc-Paper.pdf.

[35] Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:495–507, 2021.

[36] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression. *CoRR*, abs/2210.13438, 2022. doi: 10.48550/arXiv.2210.13438. URL https://doi.org/10.48550/arXiv.2210.13438.

[37] Eugene Kharitonov, Damien Vincent, Zalán Borsos, Raphaël Marinier, Sertan Girgin, Olivier Pietquin, Matt Sharifi, Marco Tagliasacchi, and Neil Zeghidour. Speak, read and prompt: High-fidelity text-to-speech with minimal supervision. *arXiv preprint arXiv:2302.03540*, 2023.

[38] Yi Ren, Xu Tan, Tao Qin, Zhou Zhao, and Tie-Yan Liu. Revisiting over-smoothness in text to speech. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8197–8213, 2022.

[39] Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, Jian Wu, Long Zhou, Shuo Ren, Yanmin Qian, Yao Qian, Michael Zeng, Xiangzhan Yu, and Furu Wei. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16:1–14, 10 2022. doi: 10.1109/JSTSP.2022.3188113.

[40] Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. Transformer vq-vae for unsupervised unit discovery and speech synthesis: Zerospeech 2020 challenge. *arXiv preprint arXiv:2005.11676*, 2020.

[41] Philip Gage. A new algorithm for data compression. *The C Users Journal archive*, 12:23–38, 1994. URL https://api.semanticscholar.org/CorpusID:59804030.

[42] RJ Skerry-Ryan, Eric Battenberg, Ying Xiao, Yuxuan Wang, Daisy Stanton, Joel Shor, Ron Weiss, Rob Clark, and Rif A Saurous. Towards end-to-end prosody transfer for expressive speech synthesis with tacotron. In *international conference on machine learning*, pages 4693–4702. PMLR, 2018.

[43] Álvaro Martín-Cortinas, Daniel Sáez-Trigueros, Iván Vallés-Pérez, Biel Tura-Vecino, Piotr Biliński, Mateusz Lajszczak, Grzegorz Beringer, Roberto Barra-Chicote, and Jaime Lorenzo-Trueba. Enhancing the stability of llm-based speech generation systems through self-supervised representations, 2024.

[44] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[45] Jeffrey De Fauw, Sander Dieleman, and Karen Simonyan. Hierarchical autoregressive image models with auxiliary decoders. *arXiv preprint arXiv:1903.04933*, 2019.

[46] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[47] Yaroslav Ganin and Victor Lempitsky. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR, 2015.

[48] Hisao Kuwabara. Acoustic properties of phonemes in continuous speech for different speaking rate. In *Proceeding of Fourth International Conference on Spoken Language Processing. ICSLP'96*, volume 4, pages 2435–2438. IEEE, 1996.

[49] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[50] Michael Hassid, Tal Remez, Tu Anh Nguyen, Itai Gat, Alexis Conneau, Felix Kreuk, Jade Copet, Alexandre Defossez, Gabriel Synnaeve, Emmanuel Dupoux, Roy Schwartz, and Yossi Adi. Textually pretrained speech language models, 2023.

[51] Won Jang, Dan Lim, Jaesam Yoon, Bongwan Kim, and Juntae Kim. Univnet: A neural vocoder with multi-resolution spectrogram discriminators for high-fidelity waveform generation, 2021.

[52] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

[53] Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. *CoRR*, abs/2010.05646, 2020. URL https://arxiv.org/abs/2010.05646.

[54] Sang-gil Lee, Wei Ping, Boris Ginsburg, Bryan Catanzaro, and Sungroh Yoon. BigVGAN: A universal neural vocoder with large-scale training. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=iTtGCMDEzS_.

[55] Eric Battenberg, R.J. Skerry-Ryan, Soroosh Mariooryad, Daisy Stanton, David Kao, Matt Shannon, and Tom Bagby. Location-relative attention mechanisms for robust long-form speech synthesis. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6194–6198, 2020. doi: 10.1109/ICASSP40776.2020.9054106.

[56] Wei Kang, Xiaoyu Yang, Zengwei Yao, Fangjun Kuang, Yifan Yang, Liyong Guo, Long Lin, and Daniel Povey. Libriheavy: a 50,000 hours asr corpus with punctuation casing and context. *arXiv preprint arXiv:2309.08105*, 2023.

[57] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2017.

[58] Weijia Ni. Sidestepping garden paths: Assessing the contributions of syntax, semantics and plausibility in resolving ambiguities. *Language and Cognitive Processes*, 11(3):283–334, 1996. doi: 10.1080/016909696387196. URL https://doi.org/10.1080/016909696387196.

[59] Hugo Quené and René Kager. The derivation of prosody for text-to-speech from prosodic sentence structure. *Computer Speech & Language*, 6(1):77–98, 1992. ISSN 0885-2308. doi: https://doi.org/10.1016/0885-2308(92)90044-5. URL https://www.sciencedirect.com/science/article/pii/0885230892900445.

[60] Edresson Casanova, Julian Weber, Christopher D Shulby, Arnaldo Candido Junior, Eren Gölge, and Moacir A Ponti. Yourtts: Towards zero-shot multi-speaker tts and zero-shot voice conversion for everyone. In *International Conference on Machine Learning*, pages 2709–2720. PMLR, 2022.

[61] Ziqiang Zhang, Long Zhou, Chengyi Wang, Sanyuan Chen, Yu Wu, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. Speak foreign languages with your own voice: Cross-lingual neural codec language modeling. *CoRR*, abs/2303.03926, 2023. doi: 10.48550/arXiv.2303.03926. URL https://doi.org/10.48550/arXiv.2303.03926.

[62] Tianrui Wang, Long Zhou, Ziqiang Zhang, Yu Wu, Shujie Liu, Yashesh Gaur, Zhuo Chen, Jinyu Li, and Furu Wei. Viola: Unified codec language models for speech recognition, synthesis, and translation. *arXiv preprint arXiv:2305.16107*, 2023.

[63] Krishna C. Puvvada, Nithin Rao Koluguri, Kunal Dhawan, Jagadeesh Balam, and Boris Ginsburg. Discrete audio representation as an alternative to mel-spectrograms for speaker and speech recognition. *CoRR*, abs/2309.10922, 2023. doi: 10.48550/arXiv.2309.10922. URL https://doi.org/10.48550/arXiv.2309.10922.

[64] Yifan Yang, Feiyu Shen, Chenpeng Du, Ziyang Ma, Kai Yu, Daniel Povey, and Xie Chen. Towards universal speech discrete tokens: A case study for asr and tts. *arXiv preprint arXiv:2309.07377*, 2023.

[65] Zhichao Huang, Chutong Meng, and Tom Ko. Repcodec: A speech representation codec for speech tokenization. *arXiv preprint arXiv:2309.00169*, 2023.

[66] Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 244–250. IEEE, 2021.

[67] Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. Speechtokenizer: Unified speech tokenizer for speech large language models. *arXiv preprint arXiv:2308.16692*, 2023.

[68] Abdelrahman Mohamed, Hung-yi Lee, Lasse Borgholt, Jakob D Havtorn, Joakim Edin, Christian Igel, Katrin Kirchhoff, Shang-Wen Li, Karen Livescu, Lars Maaløe, et al. Self-supervised speech representation learning: A review. *IEEE Journal of Selected Topics in Signal Processing*, 2022.

[69] Arnaud Joly, Marco Nicolis, Ekaterina Peterova, Alessandro Lombardi, Syed Ammar Abbas, Arent van Korlaar, Aman Hussain, Parul Sharma, Alexis Moinet, Mateusz Lajszczak, Penny Karanasou, Antonio Bonafonte, Thomas Drugman, and Elena Sokolova. Controllable emphasis with zero data for text-to-speech. *CoRR*, abs/2307.07062, 2023. doi: 10.48550/arXiv.2307.07062. URL https://doi.org/10.48550/arXiv.2307.07062.

[70] Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W. Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. In *The 9th ISCA Speech Synthesis Workshop, Sunnyvale, CA, USA, 13-15 September 2016*, page 125. ISCA, 2016. URL http://www.isca-speech.org/archive/SSW_2016/abstracts/ssw9_DS-4_van_den_Oord.html.

[71] Yunlong Jiao, Adam Gabryś, Georgi Tinchev, Bartosz Putrycz, Daniel Korzekwa, and Viacheslav Klimkov. Universal neural vocoding with parallel wavenet. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6044–6048. IEEE, 2021.

[72] Sri Karlapati, Alexis Moinet, Arnaud Joly, Viacheslav Klimkov, Daniel Sáez-Trigueros, and Thomas Drugman. Copycat: Many-to-many fine-grained prosody transfer for neural text-to-speech. In Helen Meng, Bo Xu, and Thomas Fang Zheng, editors, *Interspeech 2020, 21st Annual Conference of the International Speech Communication Association, Virtual Event, Shanghai, China, 25-29 October 2020*, pages 4387–4391. ISCA, 2020. doi: 10.21437/Interspeech.2020-1251. URL https://doi.org/10.21437/Interspeech.2020-1251.

[73] Nishant Prateek, Mateusz Lajszczak, Roberto Barra-Chicote, Thomas Drugman, Jaime Lorenzo-Trueba, Thomas Merritt, Srikanth Ronanki, and Trevor Wood. In other news: a bi-style text-to-speech model for synthesizing newscaster voice with limited data. In Anastassia Loukina, Michelle Morales, and Rohit Kumar, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 2 (Industry Papers)*, pages 205–213. Association for Computational Linguistics, 2019. doi: 10.18653/v1/n19-2026. URL https://doi.org/10.18653/v1/n19-2026.

[74] Naihan Li, Shujie Liu, Yanqing Liu, Sheng Zhao, and Ming Liu. Neural speech synthesis with transformer network. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 6706–6713, 2019.

[75] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[76] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

[77] Zihao Fu, Haoran Yang, Anthony Man-Cho So, Wai Lam, Lidong Bing, and Nigel Collier. On the effectiveness of parameter-efficient fine-tuning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 12799–12807, 2023.

[78] Sri Karlapati, Penny Karanasou, Mateusz Lajszczak, Syed Ammar Abbas, Alexis Moinet, Peter Makarov, Ray Li, Arent van Korlaar, Simon Slangen, and Thomas Drugman. Copycat2: A single model for multi-speaker TTS and many-to-many fine-grained prosody transfer. In

Hanseok Ko and John H. L. Hansen, editors, *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, pages 3363–3367. ISCA, 2022. doi: 10.21437/Interspeech.2022-367. URL https://doi.org/10.21437/Interspeech.2022-367.

[79] Detai Xin, Sharath Adavanne, Federico Ang, Ashish Kulkarni, Shinnosuke Takamichi, and Hiroshi Saruwatari. Improving speech prosody of audiobook text-to-speech synthesis with acoustic and textual contexts. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[80] Yihan Wu, Xi Wang, Shaofei Zhang, Lei He, Ruihua Song, and Jian-Yun Nie. Self-supervised context-aware style representation for expressive speech synthesis. *arXiv preprint arXiv:2206.12559*, 2022.

[81] Marcel Granero Moya, Penny Karanasou, Sri Karlapati, Bastian Schnell, Nicole Peinelt, Alexis Moinet, and Thomas Drugman. A comparative analysis of pretrained language models for text-to-speech. *CoRR*, abs/2309.01576, 2023. doi: 10.48550/arXiv.2309.01576. URL https://doi.org/10.48550/arXiv.2309.01576.

[82] Junjie Pan, Lin Wu, Xiang Yin, Pengfei Wu, Chenchang Xu, and Zejun Ma. A chapter-wise understanding system for text-to-speech in chinese novels. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6069–6073. IEEE, 2021.

[83] Rui Liu, Bin Liu, and Haizhou Li. Emotion-aware prosodic phrasing for expressive text-to-speech. *arXiv preprint arXiv:2309.11724*, 2023.

[84] Mohamed Osman. Emo-tts:parallel transformer-based text-to-speech model with emotional awareness. In *2022 5th International Conference on Computing and Informatics (ICCI)*, pages 169–174, 2022. doi: 10.1109/ICCI54321.2022.9756092.

[85] Arijit Mukherjee, Shubham Bansal, Sandeepkumar Satpal, and Rupeshkumar Mehta. Text aware emotional text-to-speech with bert. *Proc. Interspeech 2022*, pages 4601–4605, 2022.

[86] Bastian Schnell and Philip N Garner. Improving emotional tts with an emotion intensity input from unsupervised extraction. In *Proc. 11th ISCA Speech Synthesis Workshop (SSW 11)*, pages 60–65, 2021.

[87] Erik Cambria, Andrew Livingstone, and Amir Hussain. The hourglass of emotions. In *Cognitive Behavioural Systems: COST 2102 International Training School, Dresden, Germany, February 21-26, 2011, Revised Selected Papers*, pages 144–157. Springer, 2012.

[88] Goeric Huybrechts, Thomas Merritt, Giulia Comini, Bartek Perz, Raahil Shah, and Jaime Lorenzo-Trueba. Low-resource expressive text-to-speech using data augmentation. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6593–6597. IEEE, 2021.

[89] Mateusz Lajszczak, Animesh Prasad, Arent van Korlaar, Bajibabu Bollepalli, Antonio Bonafonte, Arnaud Joly, Marco Nicolis, Alexis Moinet, Thomas Drugman, Trevor Wood, and Elena Sokolova. Distribution augmentation for low-resource expressive text-to-speech. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pages 8307–8311. IEEE, 2022. doi: 10.1109/ICASSP43922.2022.9746291. URL https://doi.org/10.1109/ICASSP43922.2022.9746291.

[90] Meysam Shamsi. Tts voice corpus reduction for audio-book generation. In *6e conférence conjointe Journées d'Études sur la Parole (JEP, 33e édition), Traitement Automatique des Langues Naturelles (TALN, 27e édition), Rencontre des Étudiants Chercheurs en Informatique pour le Traitement Automatique des Langues (RÉCITAL, 22e édition). Volume 3: Rencontre des Étudiants Chercheurs en Informatique pour le TAL*, pages 193–204. ATALA; AFCP, 2020.

[91] Yihan Wu, Xu Tan, Bohan Li, Lei He, Sheng Zhao, Ruihua Song, Tao Qin, and Tie-Yan Liu. Adaspeech 4: Adaptive text to speech in zero-shot scenarios. *arXiv preprint arXiv:2204.00436*, 2022.

[92] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[93] Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, Christopher Klaiber, Pengwei Li, Daniel Licht, Jean Maillard, Alice Rakotoarison, Kaushik Ram Sadagopan, Guillaume Wenzek, Ethan Ye, Bapi Akula, Peng-Jen Chen, Naji El Hachem, Brian Ellis, Gabriel Mejia Gonzalez, Justin Haaheim, Prangthip Hansanti, Russ Howes, Bernie Huang, Min-Jae Hwang, Hirofumi Inaguma, Somya Jain, Elahe Kalbassi, Amanda Kallet, Ilia Kulikov, Janice Lam, Daniel Li, Xutai Ma, Ruslan Mavlyutov, Benjamin Peloquin, Mohamed Ramadan, Abinesh Ramakrishnan, Anna Sun, Kevin Tran, Tuan Tran, Igor Tufanov, Vish Vogeti, Carleigh Wood, Yilin Yang, Bokai Yu, Pierre Andrews, Can Balioglu, Marta R. Costa-jussà, Onur Celebi, Maha Elbayad, Cynthia Gao, Francisco Guzmán, Justine Kao, Ann Lee, Alexandre Mourachko, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Paden Tomasello, Changhan Wang, Jeff Wang, and Skyler Wang. Seamlessm4t-massively multilingual & multimodal machine translation, 2023.

[94] Sven Kachel, Adrian P. Simpson, and Melanie C. Steffens. Acoustic correlates of sexual orientation and gender-role self-concept in women's speech. *The Journal of the Acoustical Society of America*, 141(6):4793–4809, 06 2017. ISSN 0001-4966. doi: 10.1121/1.4988684. URL `https://doi.org/10.1121/1.4988684`.

[95] Sarah E. Gaither, Ariel M. Cohen-Goldberg, Calvin L. Gidney, and Keith B. Maddox. Sounding black or white: priming identity and biracial speech. *Frontiers in Psychology*, 6, 2015. ISSN 1664-1078. doi: 10.3389/fpsyg.2015.00457. URL `https://www.frontiersin.org/articles/10.3389/fpsyg.2015.00457`.

[96] Adrian Leemann, Marie-José Kolly, Francis Nolan, and Yang Li. The role of segments and prosody in the identification of a speaker's dialect. *Journal of Phonetics*, 68:69–84, 2018. ISSN 0095-4470. doi: https://doi.org/10.1016/j.wocn.2018.02.001. URL `https://www.sciencedirect.com/science/article/pii/S0095447016300365`.

A Emergent abilities testset

A.1 Questions

1. You went to the party, even though I explicitly told you not to?

2. There is another aircraft still in the air???

3. Now, seriously, you're saying I am the one to blame?

4. But she clearly doesn't want to?

5. To Hungary and back?

6. You're a copper?

7. What is Data Informed Diplomacy, with all its various manifestations?

8. What's really happening, and is there more than meets the eye?

9. How on earth is this Financial Report organized?

10. Where has Jason Abhisheki moved all the flowers to?

11. What do we do in this situation, and what are the implications for Jordan's water supply?

12. But the Brexit question remains: After all the trials and tribulations, will the ministers find the answers in time?

13. Sorry, can you restate your name and address please?

14. Here's the full story for today, would you like to learn more?

15. Mr. Chairman, your highly anticipated interview with Channel 4 has turned into a catastrophe, hasn't it?

16. Johnny boy, don't go around acting tough if you can't back it up, right?

17. Are you in favor of the Latex usage policy or you're just sucking up to leadership?

18. Is it a bird, or is it a plane?

19. Madam, have you tried turning it off and on again?

20. Were you the one with the hand-held camera or the one with a weird-looking android phone?

A.2 Emotions

1. Her hands shaking with excitement, Alice Monroe stuttered, "oh..I-I can't believe it! Is this really my acceptance letter to Harvard?" Marco cannot believe it either: "God damn it! How did you pull this off?"

2. A surge of anger flashed across the face of Matthew, as he bellowed, "You have crossed the line this time, and I won't stand for it any longer! Get out!"

3. Gazing at the panoramic view from a mountain in Iceland, Jeff Farahmand sighed deeply, whispering, "This... this is truly the face of the Divine. What more can I ask for?"

4. "You mustn't underestimate how profoundly I've missed your presence," Ito murmured, his eyes glistening with tears as he embraced his long lost sister. "You're finally back, but where do I find our lost years?"

5. "Oh my gosh! Are we really going to the Maldives? That's unbelievable!" Jennie squealed, bouncing on her toes with obvious glee.

6. "I can confidently declare that this is the most exquisite chocolate cake my taste buds have ever had the pleasure to encounter!" Mo proclaimed, savoring every bite. He could not stop eating!

7. A proud smile spread across his face as he softly said, "Son, your accomplishments fill my heart with such joy and pride." But then the smile suddenly ceased. Mike's hearts were pounding like door knocks. His dad's face now looks like that of the devil himself.

8. Choking back sobs, Mahmoud whimpered, "I simply can't fathom a life without you by my side. Don't go!"

9. His voice trembled with palpable fear as he stuttered, "There's... there's a stranger at the window. Where the hell are you all waiting for?!"

10. Tears of joy trickled down her cheeks as she yelled, "Graduating as valedictorian... this is a dream come true!"

11. Jane's eyes wide with terror, she screamed, "The brakes aren't working! What do we do now? We're completely trapped!"

12. A profound sense of realization washed over Beal as he whispered, "You've been there for me all along, haven't you? I never truly appreciated you until now."

13. Beth collapsed into his arms, sobbing uncontrollably, "I failed them, I failed them all. They're all dead! Nothing we can do will ever bring them back. How can I ever live with myself again? How?"

14. His face lit up with pure delight as he exclaimed, "We did it! We won the championship! I knew we could do it together!"

15. Overcome with guilt, Martin hung his head and muttered, "I'm so sorry. I never meant to hurt you like this. Can you ever forgive me?" It was obvious what the answer would be.

16. The queen danced around the room, eyes twinkling with mischief, "Guess what? I got the lead role in the play! Can you believe it? Well, I can't."

17. Staring into the distance, the firefighter said with a melancholic smile, "She used to sit right there, you know. I can still hear her laugh if I close my eyes." Outside the window, the rain was pouring down and gushing through every cracks.

18. The detective's voice, full of determination and fire, was heard loud and clear in the room, "No one will tell me what I can or cannot do. I'll prove them all wrong! Get me my gun. What are you all looking at me for?"

19. Overwhelmed with confusion and despair, David Darlan cried out, "What do you want from me? Why can't you just tell me what's wrong? Leave me alone!"

20. With a gentle touch and a loving smile, she reassured, "Don't worry, my love. We'll get through this together, just like we always have. I love you."

A.3 Compound Nouns

1. In the heart of Lagos, there is a public park with a serene duck pond. Nearby, the children's outdoor play area is full of joyful laughter. Nobody knows the darkness descending soon.

2. At the family reunion, my grandfather, or father-in-law for some, told many tongue-in-cheek jokes.

3. The physics teacher asked the students to build a new model solar system. Students were told to bring a tape measure and a pair of scissors, to cut the scale-model planet rings.

4. On this fateful day in 1987, the students boarded the little yellow school bus, chattering excitedly about their field trip to the zoo.

5. Hello, we are representatives from Northern Airlines. Please look out from the big second-floor window seat.

6. After years of work, Heisenberg finally published a ground-breaking cutting-edge research paper on quantum physics.

7. Recipe for a delicious breakfast sandwich: avocado, egg, and cheese on a bagel, cooked over a stovetop frying pan.

8. There is nothing more peaceful than a blue water fountain with a wooden greenhouse. Near there, Joseph installed a hard stone birdbath.

9. Prague, Czechia: Good morning, Harari! Here come the big shopping carts and last-minute video game shoppers.

10. My dog knocked over the tea table and all the books scattered across the second living room floor.

11. The hiking trail up Yahu Mountain provides a spectacular view of the sunrise. Along the path, the wooden signposts with triple-checked trail maps and green distance markers guided us.

12. The fish clock tower was striking again, reminding us of that profound changing of the guard.

13. Dean Graham sat on the packed wooden park bench, feeding the pigeons while enjoying the pleasant weather.

14. The Beckhams decided to rent a charming stone-built quaint countryside holiday cottage.

15. The construction of the new Newtown-council town hall has made huge trouble; rush-hour traffic jam has never been worse.

16. Owen Farrell has taken England to the Rugby World Cup glory, with a razor-thin-margin victory against New Zealand in France.

17. Scientists at AWS teams are making last-minute pre-launch model preparations.

18. Bad weather in Northern Europe has caused a god-awful flight check-in time of 6 AM, when even the airport food court isn't open.

19. Jake Park boasts a beautiful hand-built wooden bird feeder.

20. We visited a quaint bed-and-breakfast establishment, complete with lighthouse lamp room.

A.4 Syntactic Complexity

1. The complex houses married and single soldiers and their families.

2. Time flies like an arrow; fruit flies like a banana.

3. The rat the cat the dog chased killed ate the malt.

4. After the writer the editor the publisher hired fired quit, the company found itself in quite a bind.

5. The old man the boats on the shore were manned by had a long history of seafaring.

6. Anyone who feels that if so many more students whom we haven't actually admitted are sitting in on the course than ones we have that the room had to be changed, then probably auditors will have to be excluded, is likely to agree that the curriculum needs revision.

7. While John, who had been working late every night for a month on his novel, finally took a break to enjoy the fresh air, his neighbor, a painter who often found inspiration in the midnight moon, was just beginning her creative process.

8. In the old village with its winding roads, colorful marketplaces, a sense of history that permeates every brick, and a single traffic light, you'll find peace and simplicity.

9. The chef seasoning the fish tossed it gently.

10. As the sun dipped below the horizon, casting a golden glow over the ocean, Emily, who had spent her life dreaming of distant shores, stood on the deck of the ship, feeling a mixture of anticipation and nostalgia as her adventure began.

11. During the meeting, where Coke executives debated the future of the company, Thomas, a young intern who had discovered a solution, mustered the courage to speak, shifting the direction of the conversation, that preceded his intervention.

12. The movie that De Moya who was recently awarded the lifetime achievement award starred in 2022 was a box-office hit, despite the mixed reviews.

13. In the garden, where the flowers that the gardener who retired last year still bloomed, the children who play there every afternoon find peace and joy.

14. The scientist, Mateusz Gorka, who proposed the theory, which many experts in the field, including those who had initially been skeptical bordering on disbelieving, now support, was nominated for a prestigious award.

15. Although the meal that the chef, who had just returned from a culinary tour of Italy, prepared was delicious, the Greek guests barely noticed.

16. The book that the woman who the man who the child spoke to this morning was reading became a topic of conversation among the friends who had all read it.

17. Despite the fact that the road that led to the Five Villages, which was known for its scenic beauty, was narrow and winding, tourists flocked there throughout the year.

18. CNN journalists tracking the stories behind the officials who served during the tumultuous period when the protests rocked the nation to its core noticed significant inconsistencies in the official reports provided.

19. The musicians who performed the symphony that the composer, whose work had often been overlooked in his lifetime, wrote in his early years received a standing ovation.

20. Cars displayed in these showrooms with ENERGY-EFFICIENT AND GREEN decals prominently featured across the windshield aren't announcing environmentalism; they're virtue signaling.

A.5 Foreign Words

1. With an ample supply of joie de vivre, Mary danced through the streets of Nice, stopping only to enjoy a nice cafe with a warm croissant.

2. The modern exhibit was a mélange of styles, from German Expressionism to French Impressionism, capturing the Zeitgeist of the time.

3. As a gesture of camaraderie, the Spanish torero invited his rival, Leo the Monster, to a tapas bar, where they enjoyed jamón ibérico and the noche.

4. During Anthony's wanderlust-filled travels, he discovered the gemütlich atmosphere of many Austrian villages.

5. CloudCorp's CEO believes in gesamtkunstwerk, like integrating a symphony into a harmonious ensemble.

6. Mr. Henry, renowned for his mise en place, orchestrated a seven-course meal, each dish a pièce de résistance.

7. The fiesta, filled with música, dance, and the warmth of amigos, continued until dawn, embodying the true spirit of a Catalan celebration.

8. At the G20 Summit, leaders discussed rapprochement, trying to step away from the Schadenfreude of political rivalries.

9. After a tiring day, Sarah treated herself to a spa experience, enjoying the sauna and the jacuzzi, and relaxing with a glass of Riesling.

10. Lasso's novella, rich in allegory and imbued with a sense of ennui, drew from his experiences living in a French château up near the border.

11. The master from Osaka, Japan, dedicated himself to crafting the perfect "nigiri," with "umami" flavors dancing on the palate.

12. Mikhail Gorbachev's Reforms: Perestroika and Glasnost Define a New Era.

13. Lakshmi's yoga practice, centered around the Sanskrit concept of "ahimsa," influenced her approach to life, mirroring the teachings of Mahatma Gandhi.

14. As they strolled through the Grand Bazaar in Istanbul, they were drawn to the beautiful "kilims," the best of Turkic craftsmanship.

15. Inspired by the ancient Chinese philosophy of "Feng Shui," Li rearranged her house to create a "qi" flow throughout.

16. Embracing the Japanese aesthetic of "wabi-sabi," Hokusai's masterpieces were on full display here.

17. During Rio de Janeiro's famous Carnaval do Brasil, the streets pulsated with the rhythms of "samba".

18. The novel's protagonist, guided by the ancient Greek concept of "arete," seeks excellence and virtue, a journey reminiscent of warrior-philosopher-kings.

19. As an aficionado of Scandinavian design, Ole Gunnarsson appreciated the principle of "hygge," evident in his Danish home.

20. These soldiers - they're supposed to practice with a sense of "bushido", the samurai code of honor, but they're behaving like the imperial beasts they are.

A.6 Punctuations

1. After a moment of silence, Elena Ivanova finally spoke..., —— her words barely audible over the cracking thunder of a torrential downpour.

2. What!?! You're telling me you've never seen a single episode of 'Game of Thrones' before????! (This was not heard by Prof. Johnson, Dr. Lewis, etc.)

3. "Can anyone hear me over there??? Please, we need help!!! NOW!!!!"

4. "The Power of & and % in the Digital Age." won the first prize in this conference.

5. His latest invention (a device meant to assist in everyday chores (something he never seemed to run out of)), was nothing short of brilliant.

6. She read the label and was surprised to find — that the "natural" ingredients were actually ..... heavily processed.

7. He relayed his conversation with the bartender, saying, "I told him, 'Your 'signature' cocktail is simply a Margarita with a fancy garnish.'"

8. The presently announced laws were announced in 35°N, 80°W. Specific provisions are to be found in §12 and §17.

9. Please ensure you replace [username] and [password] with your actual credentials before logging in, like jA8!fR3$mQ1.

10. When Maria asked, 'What's happening tonight?' I replied, 'Well, John — who'll be there at 8:00 p.m. — said, "Let's meet at Sarah's place; bring games, snacks, etc., and don't be late!"'

11. "In the case of Johnson v. Smith, the court found that the defendant's actions — e.g., his failure to fulfill the terms of the contract (see sections 4.1, 4.2, and 4.3), etc. — amounted to a breach of trust."

12. When asked for his thoughts, he simply replied, «I'll be gone in 5 minutes», and left.

13. I saw Gordon listing the ingredients as follows: <tomatoes>, <fresh basil> (or dried, if unavailable - but it's essential), <olive oil>, <garlic>; salt and pepper.

14. She received an odd text from her brother: 'Emergency @ home; call ASAP! Mom & Dad are worried...#familymatters.'

15. The sign at the park's entrance stated, 'Please adhere to the following rules: no littering; no pets (except service animals); no loud music after 9 p.m.'

16. "The Art of /Slash/ and \backslash\" was the best received talk on modern internet lingo.

17. Jeb's email was brief, but to the point: 'Meeting rescheduled for 3 p.m. tomorrow – apologies for any inconvenience. Regards, J.'.

18. The Dead Sea poems contained several annotations, some of which were quite puzzling: [Section unclear]; [Translation disputed]; [Original wording lost].

19. Her travel blog post was filled with enthusiastic descriptions: 'Best trip ever!!!'; 'Amazing people & culture!'; 'Can't wait to go back...#wanderlust.'

20. He shouted, 'Everyone, please gather 'round! Here's the plan: 1) Set-up at 9:15 a.m.; 2) Lunch at 12:00 p.m. (please RSVP!); 3) Playing — e.g., games, music, etc. — from 1:15 to 4:45; and 4) Clean-up at 5 p.m.'

A.7 Paralinguistics

1. Principal Dickson began, addressing the Parkside assembly: "Ahem, I'd like to talk to you about something real serious."

2. "Aha! Now I understand," said Staff Sgt. Miller, piecing together the evidence. "The culprit left this behind. Phew."

3. "Ouch! That stings," Lilly cried, as her mother carefully applied the antiseptic. "Not beyond salvation, eh?" She dryly asked.

4. "Shh, Lucy, sshhh, we mustn't wake your baby brother," Tom whispered, as they tiptoed past the nursery.

5. "Hmm, what do you think, is it too high or two low, um... Dr. Carter?" Haim asked, handing over the instrument.

6. "Uh, well, Lisa," Tarek stuttered, nervously extending the ring he bought for god-knows how much, "mmm..will you marry me?"

7. "Yawn," Robert said, stretching out on the park bench, "this sunshine makes me sleepy."

8. "Oops! I did it again!" little Katie exclaimed, spilling her milk.

9. "Whoa, can you believe this, Mike?" Susan said, staring at the intruder. "Wow, you're right. These men ain't meanin' well."

10. James leaned back in his chair, wiped his forehead, and sighed, "Phew, haha, that was a tough meeting. Thanks for being there, Karen."

11. psst. psst. look right here.

12. "Aha! I've found it, Professor Green," exclaimed Muzi Han, holding up the rare manuscript. "This could change our entire understanding of history."

13. "Ouch, be careful, Henry!" warned his sister, as he climbed the rickety ladder.

14. David whispered to Emily as the lights dimmed in the theater, "Shh, it's starting."

15. "Hmm, I don't know about this, Jim," Mary said, looking at the folder paper. "It doesn't seem right."

16. "Uh, are you sure about this?" Tim asked nervously, looking at the steep slope before them. "Whoa, it's higher than I thought," he continued, his voice filled with trepidation. "Aha, but look at the view," Emily responded with excitement, "it's worth the climb!"

17. Ta-da! well? What do you think? This is the best right?

18. "Oops, sorry, Dad!" Jack apologized. "Ugh! you again". Dad was impatient.

19. "Whoa, what a game, Alex!" Chris exclaimed. "I've never seen anything like that final play."

20. "Phew, we made it, Martha," Tom said, collapsing into the seat after the completion of the Manhattan Project.