MAD Speech: Measures of Acoustic Diversity of Speech

Matthieu Futeral[*]
Inria Paris

Andrea Agostinelli
Google DeepMind

Marco Tagliasacchi
Google DeepMind

Neil Zeghidour
Kyutai

Eugene Kharitonov[†]
Google DeepMind

Abstract

Generative spoken language models produce speech in a wide range of voices, prosody, and recording conditions, seemingly approaching the diversity of natural speech. However, the extent to which generated speech is acoustically diverse remains unclear due to a lack of appropriate metrics. We address this gap by developing lightweight metrics of acoustic diversity, which we collectively refer to as MAD Speech. We focus on measuring five facets of acoustic diversity: voice, gender, emotion, accent, and background noise. We construct the metrics as a composition of specialized, per-facet embedding models and an aggregation function that measures diversity within the embedding space. Next, we build a series of datasets with a priori known diversity preferences for each facet. Using these datasets, we demonstrate that our proposed metrics achieve a stronger agreement with the ground-truth diversity than baselines. Finally, we showcase the applicability of our proposed metrics across several real-life evaluation scenarios. MAD Speech is made publicly available[1].

1 Introduction

Recent progress in generative spoken language modeling (Lakhotia et al., 2021; Borsos et al., 2023a; Wang et al., 2023; Kharitonov et al., 2023; Borsos et al., 2023b; Rubenstein et al., 2023; Le et al., 2023; Vyas et al., 2023; Hassid et al., 2024; Nguyen et al., 2024) brought models that can generate speech in a wide range of voices, prosody and recording conditions as found in large-scale datasets of natural speech (Kahn et al., 2020; Pratap et al., 2020; Ardila et al., 2020).

These generative models are typically evaluated in terms of faithfulness to their inputs, i.e. speaker identity and transcript (Borsos et al., 2023a; Wang et al., 2023; Kharitonov et al., 2023). Thus it remains unknown how diverse their outputs are, since those metrics neither capture variability nor take into account factors of the perceived speech diversity (i.e., recording conditions, prosody, styles, and accents). Being able to measure speech diversity would also help detect mode collapse when finetuning models (Kirk et al., 2023), selecting training & inference hyperparameters, and optimizing for human feedback (Cideron et al., 2024). Moreover, such measurements are vital when building synthetic data or mixing existing natural datasets.

In this work, we focus on measuring *acoustic diversity* of a set of speech utterances, i.e. measuring the diversity in the distribution of voices, genders, accents, emotions, and background noise. Throughout the paper, we use the term "facets" to describe these different aspects of acoustic diversity.

We propose a set of metrics, which we collectively refer to as MAD Speech (**M**easures of **A**coustic **D**iversity of Speech). We build these metrics by combining a pretrained speech representation and an aggregation function that is used to calculate the diversity of a sample in the representation space. We experiment with a set of off-the-shelf continuous speech representations such as HuBERT (Hsu et al., 2021), Wav2Vec-BERT (Chung et al., 2021a), and SoundStream (Zeghidour et al., 2021) as well as with SpeechSim, a lightweight representation model related to TRILL (Shor et al., 2020) and COLA (Saeed et al., 2021). We explore the average pairwise cosine dissimilarity and Vendi Score (Friedman and Dieng, 2022) as aggregation functions.

Further, we learn per-facet projection models with the goal of highlighting contributions of each facet "independently" from others. Such a setup is beneficial since it (a) provides a better insight into the relative contributions of different aspects of diversity, and (b) allows to avoid fusing of the dif-

ferent facets in the same metric with uncontrolled relative importance.

To evaluate the proposed metrics, we build a collection of datasets with known ground-truth acoustic diversity levels. We achieve that by subsampling standard speech datasets, while controlling for some measurable parameter that reflects the diversity of the samples along a particular facet (e.g., number of unique voices). Repeating that for all facets, we build a suite of datasets that can be used to measure to what extent a particular metric is capable of measuring acoustic diversity.

Finally, we run a set of studies to measure the impact of various improvements proposed in the literature on the acoustic diversity: fast non-autoregressive decoding of SoundStorm replacing the auto-regressive semantic-to-acoustic model of AudioLM (Borsos et al., 2023a,b) and Best-of-K decoding of SPEAR-TTS to increase quality of generated audio (Kharitonov et al., 2023). We also investigate how changes in the temperature sampling affect the acoustic diversity and compare several off-the-shelf TTS systems. These studies highlight that acoustic diversity changes non-trivially across the considered scenarios and it must be taken into account when selecting a model and a sampling setup.

2 Background

Diversity in generative models Different diversity metrics have been proposed for evaluating generative models in computer vision (Heusel et al., 2017; Naeem et al., 2020; Sajjadi et al., 2018; Kynkäänniemi et al., 2019; Jalali et al., 2023) and natural language processing (Zhang et al., 2020; Li et al., 2016; Zhu et al., 2018; Shen et al., 2019; Dušek et al., 2019; Shu et al., 2019). However these metrics are either computed on raw inputs for text (Zhu et al., 2018; Shen et al., 2019; Alihosseini et al., 2019) or based on underlying representations specific to text and images (Heusel et al., 2017; Cífka et al., 2018; Tevet and Berant, 2021) which makes them unsuitable for evaluating acoustic diversity in generative speech models.

Diversity metrics for generative audio Already the first paper proposing a generative spoken language model (GSLM) recognized the importance of measuring speech diversity (Lakhotia et al., 2021). However, since GSLM is not able to generate acoustically diverse content (e.g., it supports only a predefined set of speakers), the focus of

Lakhotia et al. (2021) was on measuring the diversity on the transcript level. In a follow-up paper, Kharitonov et al. (2022) proposed a modification of GSLM that has a more expressive prosody yet their focus was on improving and measuring within-utterance variability of prosody. To the best of our knowledge, our work is the first to study across-utterance acoustic diversity in speech generation.

Speech representations Learning compact and powerful speech representations is an extremely prolific research area (e.g., van den Oord et al., 2018; Schneider et al., 2019; Baevski et al., 2020; Saeed et al., 2021; Hsu et al., 2021; Kharitonov et al., 2021; Chung et al., 2021b; Baevski et al., 2022; Chen et al., 2022b). Often, generative audio systems are built on top of such representations by appropriately discretizing them (Lakhotia et al., 2021; Polyak et al., 2021a; Kreuk et al., 2022; Kharitonov et al., 2022; Borsos et al., 2023a; Rubenstein et al., 2023; Défossez et al., 2023). This informs our choice of representations to experiment with: HuBERT (Hsu et al., 2021) (used in the GSLM family of models (Lakhotia et al., 2021; Polyak et al., 2021b; Kreuk et al., 2021; Kharitonov et al., 2022; Nguyen et al., 2023)) and Wav2Vec-BERT (Chung et al., 2021b) (used by the AudioLM lineage (Borsos et al., 2023a; Kharitonov et al., 2023; Borsos et al., 2023b; Agostinelli et al., 2023)). We complement these off-the-shelf representations by SpeechSim, a variant of TRILL (Shor et al., 2020) and COLA (Saeed et al., 2021) that learns to contrast pairs of speech chunks. In turn, we also learn projection models that map Speech-Sim representations into embeddings that are specialized for each specific facet of diversity.

3 Learning Representations

We take a two-steps approach for measuring diversity of a set of speech utterances. First, we encode raw waveform audio in a compact representation and, afterwards, we estimate the diversity of a set of utterances in this embedding space. Hence, it is vital that the information about the relevant properties of speech is preserved in the embedding space, while the irrelevant details are omitted. We calculate embeddings in a hierarchical way: at first, we map audio to generic embeddings and then apply lightweight specialized projection models that are tailored to capture a specific facet. As a result, our approach allows us to get reliable per-facet diversity metrics. More details can be found in

Section 6.1.

SpeechSim We aim to learn speech representations where acoustically similar audio segments are close to each other in the embedding space. To do so, we train a self-supervised contrastive model with positive pairs coming from non-overlapping fixed-length chunks of the same speech utterance. The main idea being that such speech segments are likely to be acoustically similar. We refer to this model as *SpeechSim*. SpeechSim provides a speech representation that are sensitive to changes in acoustic diversity, however, it has all types of acoustic properties mixed up together. Indeed, our experiments in Section 7 show that the vanilla SpeechSim embeddings are strongly biased towards speaker voice. We therefore learn projection models (heads) on top of SpeechSim embeddings to extract different types of acoustic information.

Learning Projections We want to extract the signal relevant to each individual facet from our pre-trained embedding while suppressing information relevant to other facets. Thus, we train lightweight projection models on top of pretrained SpeechSim embeddings for each facet. To this end, we assume that we have labelled data and train these models with a contrastive objective where positive pairs are sampled from the same class w.r.t the facet.

4 From Speech Representations to Diversity Metrics

Once a set of speech utterances is encoded with one of the representations (Section 3), we can measure the diversity within the set in a variety of ways. We focus on two measures: mean pairwise dissimilarity and Vendi Score (Friedman and Dieng, 2022).

Mean pairwise dissimilarity Denoting the sample size as n and embeddings as e, we calculate:

$$1 - \frac{1}{n(n-1)} \sum_{i \neq j; i,j \leq n} \frac{e_i^T e_j}{||e_i|| \cdot ||e_j||} \quad (1)$$

We subtract the mean cosine similarity from 1, as it is more convenient to measure diversity (dissimilarity) rather than similarity. This way, an agreement with the ground-truth diversity would result in a positive correlation (see Section 5).

Vendi Score Friedman and Dieng (2022) proposed to measure diversity of a sample as the exponential of the Shannon entropy of the eigenvalues of the pairwise similarity matrix, normalized by the number of vectors. In practice, we built it by

computing pairwise cosine similarity and dividing by the number of examples. Formally, denoting the eigenvalues of the normalized similarity matrix as $\{\lambda_k\}_{k=1}^n$, Vendi Score is calculated as:

$$\exp\left[-\sum_{i=1}^n \lambda_i \log \lambda_i\right] \quad (2)$$

Eq. 2 is well-defined, as the normalized similarity matrix is positive semi-definite and $0 \leq \lambda_i \leq 1$.

5 Evaluating Diversity Metrics

5.1 Our Approach

We break down the acoustic diversity into the following facets: speaker voice, gender, emotion (or style), accent, and background noise. For each of these facets, we repeatedly sample subsets from a large dataset of natural speech utterances while systematically increasing the level of diversity in a controllable way, as further described below. Next, we calculate Spearman's rank correlation (Spearman, 1904) between diversity metric scores and the ground-truth diversity level's rank. Metrics with a higher rank correlation are better at representing a particular facet of the acoustic diversity.

In the following, we describe the general template for building a series of datasets of controllable diversity, using voice diversity as a reference. We fix the sample size *a priori* and define voice diversity as the number of distinct speakers in a set of examples. To build one dataset, we sample examples from a natural dataset such that we obtain the required number of speakers and have every speaker represented by the same number of utterances. We create a series of datasets by varying the number of distinct speakers. We also ensure that other facets do not differ from one dataset to another as much as possible so that we make sure we indeed test speaker voice diversity.

This process is similar for other facets. We determine gender diversity as a proportion of female voices; background noise diversity is characterized by the number of different background noise classes; and emotion (accent) diversity is determined by computing the entropy of the distribution of the different emotion (accent) classes.

5.2 Evaluation data

For each facet and each diversity level, we randomly sample 100 sets using different seeds.

Speakers voices We build evaluation (validation) sets from LibriTTS (Zen et al., 2019) test-clean & test-other (dev-clean & dev-other). Each set contains 200 utterances of a single gender (either all male voices or all female voices) to test whether a particular metric is sensitive to the voice diversity *independently* from the gender diversity. We vary the number of distinct voices in $\{5, 10, 15, 20, 25, 33\}$.

To test the influence of gender diversity for the metrics that target voice diversity, we generate additional data where the voice and the gender diversity are pushed in the opposite directions: having a low number of speakers with uniform gender distribution, and a single gender (selected randomly) when the number of speakers is high.

Gender We build evaluation (validation) sets from LibriTTS (Zen et al., 2019) test-clean & test-other (dev-clean & dev-other). Each set contains 100 utterances. We change the ratio of female voices from 0.0 to 1.0 in increments of 0.1.

We built two types of evaluation samples, one with equal number of different speaker voices in the samples and another one where number of different speaker voices is maximum (resp. minimum) when there is few (resp. lots of) gender variation to test if the gender metric is speaker-sensitive.

Emotion We build evaluation (validation) sets from 40% (10%) of EmoV (Adigwe et al., 2018) randomly sampled and test (dev) splits of Expresso (Nguyen et al., 2023). We built two types of samples: (1) with equal number of different speaker voices and (2) with a single speaker voice for datasets with high entropy but all speaker voices—same number of utterances per speaker—for datasets with low entropy. Speakers in test and dev sets are identical to those found in the training set. Emotion classes match in train, dev, and test.

Accent We build evaluation (validation) sets from test (dev) sets of VCTK (Veaux et al., 2016). Accent classes are common between train, dev, and tests sets. Again, we built two types of samples (1) with equal number of different speaker voices and (2) with the minimum number of different speaker voices for high-entropy datasets but all speaker voices— equally represented—for low-entropy datasets. Speakers in test and dev sets are distinct from those found in the training set.

Background Noise We build evaluation and validation sets by sampling 20% and 10% of AudioSet,

respectively (Gemmeke et al., 2017). We select examples that have exactly two tags, one of them being "Speech". We treat the second tag as a class of the background noise. Next, we construct a sample by randomly selecting 100 examples with varying number of distinct noise classes (1, 5, 10, 25, 50, or 100).

6 Experimental Setup

Details about the training data and the baseline representations we use can be found in Appendix A and Appendix B respectively.

6.1 Training details

SpeechSim The input to SpeechSim is a 6-seconds chunk of audio with sampling rate of 16kHz. We compute the mel-spectrogram using a window length of 512, hop length of 256 and 156 bins. This produces 376 temporal frames of dimension 156. The model is based on a small ViT architecture (Dosovitskiy et al., 2020), containing 12 layers with each layer having 6 attention heads, embedding dimension of 512, and a feed-forward layer of dimension 1024 (25M parameters). The final embeddings are averaged across the time axis and are projected into a 192-dimensional vector. We train SpeechSim with the semi-hard triplet loss (Schroff et al., 2015), using a total batch size of 3840. We train the model for a total of 10^5 steps.

Projection models We train projection models on top of SpeechSim embeddings with. a standard contrastive loss (Radford et al., 2021). The positive pairs are constructed by taking two examples with the same labels (e.g., same emotion). Except for the gender projector, all models have 2 fully connected layer (of size 256 and 128) with a GELU activation (Hendrycks and Gimpel, 2016) in-between. In the case of the gender projector, we found that a 4-layers network works better (dimensions of 256, 256, 128 and 128). Dropout rate of 0.1 is applied to SpeechSim input representations. We trained the projection models using Adam (Kingma and Ba, 2014) and use a learning rate of 1e-4, and a batch size of 128. We also apply a weight decay of 1e-3 for projection models related to speaker voices, emotion and background noise and 1e-4 for other facets. We select hyperparameters using validation loss.

All embeddings that we use in our experiments have a single vector representing the entire utterance. To achieve that, we average embeddings over

	Male		Female	
	Avg. Cosine	Vendi score	Avg. Cosine	Vendi score
SoundStream	0.544 (± 0.312)	0.759 (± 0.175)	0.234 (± 0.426)	0.731 (± 0.284)
Wav2vec-BERT	-0.448 (± 0.194)	-0.419 (± 0.193)	-0.118 (± 0.258)	-0.094 (± 0.245)
HuBERT	-0.575 (± 0.094)	-0.447 (± 0.202)	-0.178 (± 0.291)	0.007 (± 0.321)
SpeechSim	0.811 (± 0.074)	**1.000** (± 0.000)	0.834 (± 0.073)	**1.000** (± 0.000)
Trained from scratch	0.480 (± 0.200)	0.962 (± 0.044)	0.666 (± 0.200)	<u>0.963</u> (± 0.053)
SpeechSim/Voice	0.823 (± 0.063)	<u>0.993</u> (± 0.019)	0.881 (± 0.077)	**1.000** (± 0.000)

Table 1: Voice diversity. Average Spearman correlations (± standard error) between number of distinct speakers and diversity scores induced by speech representations. Male (Female) refer to samples with male-only (female-only) voices. Best results are in bold, second best results are underlined.

Female voices ratio	≤ 0.5		> 0.5	
	Avg. Cosine	Vendi score	Avg. Cosine	Vendi score
SoundStream	0.095 (± 0.463)	0.092 (± 0.435)	0.117 (± 0.410)	0.097 (± 0.367)
Wav2vec-BERT	-0.035 (± 0.434)	0.026 (± 0.427)	0.159 (± 0.406)	0.105 (± 0.415)
HuBERT	0.117 (± 0.431)	0.223 (± 0.448)	0.109 (± 0.445)	0.004 (± 0.444)
SpeechSim	0.186 (± 0.473)	0.132 (± 0.440)	0.129 (± 0.438)	0.162 (± 0.455)
Trained from scratch	<u>0.861</u> (± 0.076)	0.478 (± 0.320)	<u>0.912</u> (± 0.047)	0.538 (± 0.317)
SpeechSim/Gender	**0.902** (± 0.108)	0.708 (± 0.249)	**0.946** (± 0.057)	0.742 (± 0.216)

Table 2: Gender diversity. Average Spearman correlations between proportion of female voices and diversity scores induced by speech representations.

the time axis. We found this to perform better than max-pooling.

7 Evaluating MAD Speech

In this Section, we use the collection of per-facet diversity benchmarks (Section 5) to assess which speech representations are more suitable for estimating diversity of a set of speech utterances when only one facet is changed (Section 7.1). Next, we run a series of experiments with two facets changed in the opposite directions (Section 7.2).

7.1 Variability in a single facet

Voice Diversity From Table 1 we firstly notice that, across all columns, SpeechSim shows higher correlation scores than off-the-shelf embedding models and the models trained from scratch. Further voice specialization of SpeechSim (Speech-Sim/Voice) results in similar scores. Finally, on comparing average cosine vs. Vendi Score aggregation, we observe that often Vendi Score shows higher correlations.

Gender diversity In Table 2 we report correlation scores for the gender diversity. Again, we split in two groups: male-voice and female voice-dominant. From the results, we notice that the correlations showed by the non-specialized embed-

dings are extremely weak. However, both specialized models reach very high correlations (e.g., up to 0.946 for SpeechSim/Gender). All in all, gender-specific projector gets the higher correlation score. In this setup, Vendi Score underperforms w.r.t. the average cosine dissimilarity.

Emotion diversity In Table 3 we report average Spearman correlations for the tested representation models, when tested on EmoV and Expresso separately. From these results, we see that, generally, SpeechSim performs better than other generic representations across all configurations. Equally, the specialized projection of SpeechSim for the emotion facet, SpeechSim/Emotion performs best across all representations. As with voice diversity, Vendi Score gets higher correlations than average cosine dissimilarity.

To make sure that our models can measure diversity of emotions beyond that is covered by the (limited) label sets of EmoV and Expresso, we run an experiment where the projection model is trained on one dataset and tested on another (the only common label is "neutral"). Table 3 shows that SpeechSim finetuned on Expresso does transfer to EmoV classes as it obtains a high Spearman correlation. However, SpeechSim finetuned on EmoV gets a Spearman correlation equal to 0.321. Digging

	EmoV		Expresso	
	Avg. Cosine	Vendi score	Avg. Cosine	Vendi score
SoundStream	0.137 (± 0.573)	0.714 (± 0.250)	0.357 (± 0.383)	0.495 (± 0.348)
Wav2vec-BERT	0.682 (± 0.182)	0.594 (± 0.080)	0.668 (± 0.170)	0.705 (± 0.070)
HuBERT	0.785 (± 0.148)	0.683 (± 0.088)	0.725 (± 0.154)	0.725 (± 0.064)
SpeechSim	0.809 (± 0.200)	0.823 (± 0.193)	0.739 (± 0.249)	0.864 (± 0.128)
Trained from scratch	0.991 (± 0.020)	0.962 (± 0.051)	<u>0.998</u> (± 0.008)	0.988 (± 0.018)
Trained from scratch/Expresso	0.498 (± 0.250)	0.513 (± 0.274)	-	-
Trained from scratch/EmoV	-	-	0.247 (± 0.562)	0.251 (± 0.512)
SpeechSim/Emotion	**0.995** (± 0.016)	<u>0.993</u> (± 0.019)	**0.999** (± 0.007)	0.987 (± 0.017)
SpeechSim/Emotion-Expresso	0.641 (± 0.279)	0.709 (± 0.243)	-	-
SpeechSim/Emotion-EmoV	-	-	0.273 (± 0.465)	0.321 (± 0.464)

Table 3: Emotion diversity. Average Spearman correlations between the classes entropy in EmoV and Expresso and diversity scores induced by the speech representations. SpeechSim/Emotion-Expresso (resp. Speech/Emotion-EmoV) refers to SpeechSim emotion head trained on Expresso (resp. EmoV).

	Avg. Cosine	Vendi score
SoundStream	0.234 (± 0.517)	0.217 (± 0.581)
Wav2vec-BERT	0.425 (± 0.415)	0.557 (± 0.376)
HuBERT	0.372 (± 0.425)	0.526 (± 0.366)
SpeechSim	0.446 (± 0.476)	0.574 (± 0.328)
Trained from scratch	<u>0.995</u> (± 0.016)	**1.000** (± 0.000)
SpeechSim/Accent	0.991 (± 0.021)	**1.000** (± 0.000)

Table 4: Accent diversity. Average Spearman correlations between the accent class entropy and diversity scores induced by speech representations.

	Avg. Cosine	Vendi score
SoundStream	0.141 (± 0.390)	0.265 (± 0.433)
Wav2vec-BERT	-0.014 (± 0.417)	-0.070 (± 0.452)
HuBERT	0.218 (± 0.356)	**0.531** (± 0.332)
SpeechSim	0.254 (± 0.309)	0.466 (± 0.322)
Trained from scratch	0.173 (± 0.355)	0.355 (± 0.404)
SpeechSim/Noise	0.116 (± 0.359)	<u>0.469</u> (± 0.413)

Table 5: Background noise diversity. Average Spearman correlations between the number of different classes of noise and diversity scores induced by speech representations.

deeper, we additionally compute the correlation of SpeechSim/Emotion-EmoV scores where we average all scores (obtained with different seeds) so that we obtain a single score for each diversity level. In this, SpeechSim/Emotion-Emov gets a correlation of 0.964 which means that SpeechSim finetuned on EmoV generalizes well on average.

Accent diversity We report average correlations in Table 4. Here, again, we notice that SpeechSim has higher correlations than other non-specialized embedding models. Both specialized model perform similarly good with perfect correlations. As before, Vendi Score leads to better correlations across all setups.

Background noise diversity We report our results in Table 5. We see that generally all representations have somewhat low levels of correlation and high variability across seeds probably due to the fact that AudioSet is pretty noisy and multiple labels can be found in a single example. That said, SpeechSim/Noise combined with Vendi Score obtains the second highest correlation, following HuBERT.

We additionally compute the correlation of the

SpeechSim/Noise scores where we average all scores (obtained with different seeds) so that we obtain a single score for each diversity level. In this setup, SpeechSim/Noise gets a Spearman correlation of 1.00, indicating that it works very well on average.

All in all, Tables 1 to 5 show that SpeechSim generic representations achieve higher correlation scores than off-the-shelf embedding models while having lower latency, this motivates us to rely on it as a basis for per-facet models.

7.2 Are facets mixed within metrics?

Above we assessed how good the proposed scores are in detecting changes in a single factor of diversity with the rest fixed. However, one can expect that in a real-life scenario many factors can change simultaneously. In this Section, we verify whether this behavior takes place.

Gender and Voices We design a sequence of samples such that the voice and gender diversity are changing in the opposite directions. We continuously change the proportion of female voices

	Avg. cosine	Vendi score
< 0.5 female voices		
Speech Sim	-0.709 (± 0.108)	-0.734 (± 0.107)
SpeechSim/Speaker	-0.657 (± 0.088)	-0.686 (± 0.092)
SpeechSim/Gender	0.781 (± 0.259)	0.147 (± 0.341)
≥ 0.5 female voices		
SpeechSim	-0.706 (± 0.110)	-0.709 (± 0.120)
SpeechSim/Speaker	-0.648 (± 0.080)	-0.668 (± 0.092)
SpeechSim/Gender	0.827 (± 0.285)	0.313 (± 0.342)

Table 6: Changing voice and gender diversity in the opposite directions. Average Spearman correlation between the gender diversity and the diversity scores induced by speech representations.

(a) SpeechSim/Gender, avg. pairwise dissim.
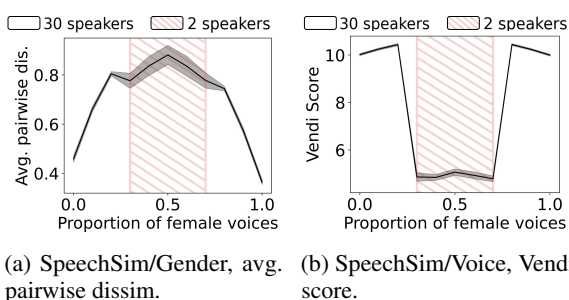(b) SpeechSim/Voice, Vendi score.

Figure 1: Changing the proportion of female voices and number of distinct speaker voices simultaneously. Grey area represents 95% CI.

from 0.0 to 1.0: the gender diversity grows at first, peaks when the ratio is 0.5, and then starts to decrease. While the gender diversity increases, we reduce the number of speakers (from 30 to 2) and when the gender diversity starts to fall, we increase the number of speakers back (from 2 to 30).

We report results in Table 6. From this results we see that SpeechSim/Gender is positively correlated with the gender diversity. In contrast, the SpeechSim/Voice metric tracks the speaker diversity and hence is negatively correlated with the gender diversity. Finally, the original SpeechSim embeddings have negative correlation with the gender diversity. Moreover, its correlation coefficient is smaller than that of SpeechSim/Voice. From that we conclude that the vanilla SpeechSim measure overemphasizes voice over gender diversity.

We further illustrate this experiment in Figure 1, reporting SpeechSim/Gender and SpeechSim/Voice scores as the female voice ratio changes. From Figure 1 we see that SpeechSim/Gender is almost invariant to the changes in the number of speakers. Conversely, SpeechSim/Voice is relatively insensitive to changes of gender diversity.

We run a similar experiment for accents and emotions and report it in Appendix D. We observe similar findings, raw SpeechSim is highly influenced by Speaker voice variations while specialized facets are unaffected by changes in voices.

Overall, the experiments reported in this Section and Appendix D support the necessity of introducing per-facet projection models. While SpeechSim embeddings might provide reasonably good correlations with the ground-truth diversity when only one facet is changed, they implicitly mix all facets into a single score. We find this inconvenient, as (a) changes in one facet can mask changes in another, (b) the user might want to control the trade-off explicitly. Hence, per-facet models are preferrable.

8 Comparing Generative Models

In this section, we show that despite that the acoustic diversity was mostly ignored in the literature when proposing new models and approaches, it does change.

We start by assessing whether the introduction of SoundStorm (Borsos et al., 2023b) to the AudioLM framework (Borsos et al., 2023a) lead to any hidden penalties in acoustic diversity (Section 8.1) before comparing acoustic diversity of off-the-shelf TTS models (Section 8.2). Additional studies on the effect of the Best-K sampling technique used by Kharitonov et al. (2023) on acoustic diversity and how temperature changes acoustic diversity of the synthesized speech can be found in Appendix C.1 and Appendix C.2. In this Section, we use the same terminology as Borsos et al. (2023a) where semantic tokens refer to tokens learned with a self-supervised masked language modeling objective (Chung et al., 2021a; Hsu et al., 2021) while acoustic tokens refer to tokens learned with a full reconstruction objective from a residual vector quantizer (Zeghidour et al., 2021).

8.1 Semantic-to-Acoustic Token Conversion: SoundStorm vs. AudioLM

Generating acoustic tokens from semantic tokens is a core part of AudioLM (Borsos et al., 2023a). This step is believed to be responsible for modelling the acoustic details of the generated audio, such as voice, background noise, etc. Borsos et al. (2023b) proposed to implement it with a non-autoregressive Transformer, yielding the SoundStorm model that dramatically improved inference time. However, it is unclear whether this change harms acoustic

Voice	Gender	Emotion	Accent	Background noise
0.0	62.5	6.3	1.6	4.7

Table 7: Semantic-to-acoustic tokens resynthesis task. The proportion of sets of utterances (%) where the AudioLM stage 2 model generates samples that are more diverse than those generated by SoundStorm.

	Voice	Gender	Emotion	Accent	Back. Noise
Bark TTS	39.41	0.90	8.29	8.30	3.84
Tortoise TTS	30.94	0.92	8.73	8.17	3.17
StyleTTS 2	31.54	0.37	7.73	6.81	2.60
FastSpeech 2	19.33	0.29	6.74	5.63	2.42

Table 8: Acoustic diversity scores for off-the-shelf TTS systems (higher is more diverse). We report Vendi score for all facets except for gender, which uses cosine dissimilarity.

diversity of the generated audio.

We take 32 male- and 32 female-voice utterances from hold-out LibriTTS test-clean and represent them as sequences of semantic tokens. Next, we convert these semantic token sequences to acoustic token sequences using both models, repeating the process 128 times. We convert the acoustic tokens into audio using the corresponding Sound-Stream codec (Zeghidour et al., 2021). Finally, for each combination of a source utterance and a model, we calculate the diversity of the produced audio samples. We use exact same SoundStorm and AudioLM models as Borsos et al. (2023b).

In Table 7 we report, for each facet of the acoustic diversity, the ratio of utterances where AudioLM generated higher-diversity audio than SoundStorm. We see that SoundStorm produces audio that is more diverse in voices, emotions, accents, and background noise, while being behind in the gender diversity.

8.2 Acoustic diversity in Text-to-Speech systems

Next, we measure acoustic diversity of several publicly available TTS systems. Here, our goal is to show that MAD Speech brings a new dimension to the evaluation of real-world TTS systems. We compare *Bark TTS*[2], *StyleTTS 2* (Li et al., 2024), *Tortoise TTS* (Betker, 2023a) and *FastSpeech 2* (Ren et al., 2020). Details about these models can be found in Appendix E. As some of these systems are deterministic, we adopt a different setup and synthesize all transcripts in shuffled LibriTTS test-clean and calculate the average acoustic diversity across batches of audio of size 128. We report the results in Table 8.

We see that Bark and Tortoise TTS are the most diverse across all facets, likely due to training on more diverse internet data. Interestingly, Tortoise TTS tends to be more diverse than Bark in emotions while less or equally diverse in other facets. StyleTTS 2 is considerably less diverse in the gen-

der facet — just as one would expect, given that the generation is restricted to female voices. Fast-Speech 2 is the least acoustically diverse mainly due to being trained on a very homogeneous dataset.

All in all, the comparisons we carried out in this Section and Appendix C show that previously proposed changes lead to non-trivial and sometimes even non-monotonic changes in per-facet acoustic diversity. Equally, the sampling temperature also affects the resulting acoustic diversity. We also found that MAD Speech can be useful in detecting acoustic diversity trends in off-the-shelf TTS systems. Overall, we believe those findings highlight the necessity for including acoustic diversity metrics in the standard evaluations.

9 Conclusions & Future work

We introduced MAD Speech, a set of metrics for evaluating acoustic diversity in speech. MAD Speech metrics are calculated in two steps: (1) a set of utterances is mapped to an embedding space, (2) acoustic diversity is estimated by comparing the resulting embeddings with an aggregation function (cosine dissimilarity or Vendi Score). Calculating embeddings is done by firstly applying a generic representation model with a subsequent metric-specific projection. We focus on five facets of acoustic diversity: voices, gender, emotion, accents, and background noise.

In order to validate the proposed metrics, we built a collection of datasets with a controlled level of acoustic diversity. Having these sets allows us to evaluate a metric by calculating its Spearman rank correlation with the ground-truth acoustic diversity.

Our empirical study demonstrated that our proposed metrics have a highest agreement with the ground-truth diversity levels when compared to baseline approaches and that they are insensitive to variations of other facets of diversity.

Finally, we highlighted a necessity of controlling

---

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

acoustic diversity when developing new models and approaches. To this end, we took some recent modelling improvements and demonstrated that they, in fact, affect acoustic diversity of speech.

Limitations

By building MAD Speech, we make it possible to holistically assess progress in generative speech and reveal possible hidden biases in the acoustic scenes. Yet, our metrics are machine-learned models themselves and are limited by the nature of the data we used to build them. Specifically, the public datasets we use are limited to English language. The labels in the data can be noisy and only represent a limited reflection of the world. We also made the choice to treat gender as a binary variable to be consistent with existing datasets, this can lead to minor measurement inaccuracies when facing ambiguously-gendered voices (Stoidis and Cavallaro, 2022).

References

Adaeze Adigwe, Noé Tits, Kevin El Haddad, Sarah Ostadabbas, and Thierry Dutoit. 2018. The emotional voices database: Towards controlling the emotion dimension in voice generation systems. *arXiv preprint arXiv:1806.09514*.

Andrea Agostinelli, Timo I Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, et al. 2023. MusicLM: Generating music from text. *arXiv preprint arXiv:2301.11325*.

Danial Alihosseini, Ehsan Montahaei, and Mahdieh Soleymani Baghshah. 2019. Jointly measuring diversity and quality in text generation models. In *Proceedings of the Workshop on Methods for Optimizing and Evaluating Neural Language Generation*, pages 90–98, Minneapolis, Minnesota. Association for Computational Linguistics.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. 2022. data2vec: A general framework for self-supervised learning in speech, vision and language. *Preprint*, arXiv:2202.03555.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Preprint*, arXiv:2006.11477.

Evelina Bakhturina, Vitaly Lavrukhin, Boris Ginsburg, and Yang Zhang. 2021. Hi-fi multi-speaker english tts dataset. *arXiv preprint arXiv:2104.01497*.

James Betker. 2023a. Better speech synthesis through scaling. *arXiv preprint arXiv:2305.07243*.

James Betker. 2023b. Better speech synthesis through scaling. *Preprint*, arXiv:2305.07243.

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

Zalán Borsos, Matt Sharifi, Damien Vincent, Eugene Kharitonov, Neil Zeghidour, and Marco Tagliasacchi. 2023b. Soundstorm: Efficient parallel audio generation. *arXiv preprint arXiv:2305.09636*.

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, Jian Wu, Michael Zeng, Xiangzhan Yu, and Furu Wei. 2022a. WavLM: Large-scale self-supervised pre-training for full stack speech processing. *IEEE JSTSP*, 16(6).

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2022b. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1505–1518.

Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. 2021a. 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.

Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu. 2021b. W2V-Bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. In *ASRU*.

Geoffrey Cideron, Sertan Girgin, Mauro Verzetti, Damien Vincent, Matej Kastelic, Zalán Borsos, Brian McWilliams, Victor Ungureanu, Olivier Bachem, Olivier Pietquin, Matthieu Geist, Léonard Hussenot, Neil Zeghidour, and Andrea Agostinelli. 2024. Musicrl: Aligning music generation to human preferences. *Preprint*, arXiv:2402.04229.

Ondřej Cífka, Aliaksei Severyn, Enrique Alfonseca, and Katja Filippova. 2018. Eval all, trust a few, do wrong to none: Comparing sentence generation models. *arXiv preprint arXiv:1804.07972*.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2023. High fidelity neural audio compression. *Transactions on Machine Learning Research*. Featured Certification, Reproducibility Certification.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. 2019. Evaluating the state-of-the-art of end-to-end natural language generation: The e2e nlg challenge. *Computer Speech & Language*, 59.

Dan Friedman and Adji Bousso Dieng. 2022. The vendi score: A diversity evaluation metric for machine learning. *arXiv preprint arXiv:2210.02410*.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780.

Michael Hassid, Tal Remez, Tu Anh Nguyen, Itai Gat, Alexis Conneau, Felix Kreuk, Jade Copet, Alexandre Defossez, Gabriel Synnaeve, Emmanuel Dupoux, et al. 2024. Textually pretrained speech language models. *Advances in Neural Information Processing Systems*, 36.

Dan Hendrycks and Kevin Gimpel. 2016. Gaussian error linear units (gelus). *arXiv preprint arXiv:1606.08415*.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3451–3460.

Keith Ito and Linda Johnson. 2017. The LJ speech dataset. https://keithito.com/LJ-Speech-Dataset/.

Mohammad Jalali, Cheuk Ting Li, and Farzan Farnia. 2023. An information-theoretic evaluation of generative models in learning multi-modal distributions. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Jacob Kahn, Morgane Rivière, Weiyi Zheng, Evgeny Kharitonov, Qiantong Xu, Pierre-Emmanuel Mazaré, Julien Karadayi, Vitaliy Liptchinsky, Ronan Collobert, Christian Fuegen, et al. 2020. Libri-Light: A benchmark for ASR with limited or no supervision. In *IEEE ICASSP*.

Eugene Kharitonov, Ann Lee, Adam Polyak, Yossi Adi, Jade Copet, Kushal Lakhotia, Tu Anh Nguyen, Morgane Riviere, Abdelrahman Mohamed, Emmanuel Dupoux, and Wei-Ning Hsu. 2022. Text-free prosody-aware generative spoken language modeling. In *ACL*.

Eugene Kharitonov, Morgane Rivière, Gabriel Synnaeve, Lior Wolf, Pierre-Emmanuel Mazaré, Matthijs Douze, and Emmanuel Dupoux. 2021. Data augmenting contrastive learning of speech representations in the time domain. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 215–222.

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

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. 2023. Understanding the effects of rlhf on llm generalisation and diversity. *arXiv preprint arXiv:2310.06452*.

Felix Kreuk, Adam Polyak, Jade Copet, Eugene Kharitonov, Tu-Anh Nguyen, Morgane Rivière, Wei-Ning Hsu, Abdelrahman Mohamed, Emmanuel Dupoux, and Yossi Adi. 2021. Textless speech emotion conversion using decomposed and discrete representations. *arXiv preprint arXiv:2111.07402*.

Felix Kreuk, Gabriel Synnaeve, Adam Polyak, Uriel Singer, Alexandre Défossez, Jade Copet, Devi Parikh, Yaniv Taigman, and Yossi Adi. 2022. Audiogen: Textually guided audio generation. *arXiv preprint arXiv:2209.15352*.

Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. 2019. Improved precision and recall metric for assessing generative models. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Kushal Lakhotia, Eugene Kharitonov, Wei-Ning Hsu, Yossi Adi, Adam Polyak, Benjamin Bolte, Tu-Anh

Nguyen, Jade Copet, Alexei Baevski, Abdelrahman Mohamed, et al. 2021. On generative spoken language modeling from raw audio. *TACL*.

Matthew Le, Apoorv Vyas, Bowen Shi, Brian Karrer, Leda Sari, Rashel Moritz, Mary Williamson, Vimal Manohar, Yossi Adi, Jay Mahadeokar, and Wei-Ning Hsu. 2023. Voicebox: Text-guided multilingual universal speech generation at scale. In *Advances in Neural Information Processing Systems*, volume 36, pages 14005–14034. Curran Associates, Inc.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Yinghao Aaron Li, Cong Han, Vinay Raghavan, Gavin Mischler, and Nima Mesgarani. 2024. Styletts 2: Towards human-level text-to-speech through style diffusion and adversarial training with large speech language models. *Advances in Neural Information Processing Systems*, 36.

Muhammad Ferjad Naeem, Seong Joon Oh, Youngjung Uh, Yunjey Choi, and Jaejun Yoo. 2020. Reliable fidelity and diversity metrics for generative models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 7176–7185. PMLR.

Tu Anh Nguyen, Wei-Ning Hsu, Antony d'Avirro, Bowen Shi, Itai Gat, Maryam Fazel-Zarani, Tal Remez, Jade Copet, Gabriel Synnaeve, Michael Hassid, et al. 2023. Expresso: A benchmark and analysis of discrete expressive speech resynthesis. *arXiv preprint arXiv:2308.05725*.

Tu Anh Nguyen, Benjamin Muller, Bokai Yu, Marta R Costa-Jussa, Maha Elbayad, Sravya Popuri, Paul-Ambroise Duquenne, Robin Algayres, Ruslan Mavlyutov, Itai Gat, et al. 2024. Spirit-lm: Interleaved spoken and written language model. *arXiv preprint arXiv:2402.05755*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. LibriSpeech: an ASR corpus based on public domain audio books. In *ICASSP*.

Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux. 2021a. Speech Resynthesis from Discrete Disentangled Self-Supervised Representations. In *Proc. Interspeech 2021*.

Adam Polyak, Yossi Adi, Jade Copet, Eugene Kharitonov, Kushal Lakhotia, Wei-Ning Hsu, Abdelrahman Mohamed, and Emmanuel Dupoux. 2021b. Speech Resynthesis from Discrete Disentangled Self-Supervised Representations. In *Interspeech*.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. MLS: A large-scale multilingual dataset for speech research. *arXiv preprint arXiv:2012.03411*.

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

Yi Ren, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao, and Tie-Yan Liu. 2020. FastSpeech 2: Fast and high-quality end-to-end text to speech. In *ICLR*.

Paul K. Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, Hannah Muckenhirn, Dirk Padfield, James Qin, Danny Rozenberg, Tara Sainath, Johan Schalkwyk, Matt Sharifi, Michelle Tadmor, Ramanovich, Marco Tagliasacchi, Alexandru Tudor, Mihajlo Velimirović, Damien Vincent, Jiahui Yu, Yongqiang Wang, Vicky Zayats, Neil Zeghidour, Yu Zhang, Zhishuai Zhang, Lukas Zilka, and Christian Frank. 2023. AudioPaLM: A large language model that can speak and listen.

Aaqib Saeed, David Grangier, and Neil Zeghidour. 2021. Contrastive learning of general-purpose audio representations. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3875–3879.

Mehdi S. M. Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. 2018. Assessing generative models via precision and recall. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Steffen Schneider, Alexei Baevski, Ronan Collobert, and Michael Auli. 2019. wav2vec: Unsupervised Pre-Training for Speech Recognition. In *Proc. Interspeech 2019*, pages 3465–3469.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823.

Tianxiao Shen, Myle Ott, Michael Auli, and Marc'Aurelio Ranzato. 2019. Mixture models for diverse machine translation: Tricks of the trade. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 5719–5728. PMLR.

Joel Shor, Aren Jansen, Ronnie Maor, Oran Lang, Omry Tuval, Felix de Chaumont Quitry, Marco

Tagliasacchi, Ira Shavitt, Dotan Emanuel, and Yinnon Haviv. 2020. Towards learning a universal non-semantic representation of speech. *arXiv preprint arXiv:2002.12764*.

Raphael Shu, Hideki Nakayama, and Kyunghyun Cho. 2019. Generating diverse translations with sentence codes. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1823–1827, Florence, Italy. Association for Computational Linguistics.

C. Spearman. 1904. The proof and measurement of association between two things. *American Journal of Psychology*, 15:88–103.

Dimitrios Stoidis and Andrea Cavallaro. 2022. Generating gender-ambiguous voices for privacy-preserving speech recognition. In *Interspeech 2022*, pages 4237–4241.

Guy Tevet and Jonathan Berant. 2021. Evaluating the evaluation of diversity in natural language generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 326–346, Online. Association for Computational Linguistics.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv:1807.03748*.

Christophe Veaux, Junichi Yamagishi, and Kirsten MacDonald. 2016. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit.

Apoorv Vyas, Bowen Shi, Matthew Le, Andros Tjandra, Yi-Chiao Wu, Baishan Guo, Jiemin Zhang, Xinyue Zhang, Robert Adkins, William Ngan, Jeff Wang, Ivan Cruz, Bapi Akula, Akinniyi Akinyemi, Brian Ellis, Rashel Moritz, Yael Yungster, Alice Rakotoarison, Liang Tan, Chris Summers, Carleigh Wood, Joshua Lane, Mary Williamson, and Wei-Ning Hsu. 2023. Audiobox: Unified audio generation with natural language prompts. *Preprint*, arXiv:2312.15821.

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

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. 2021. SoundStream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. Libritts: A corpus derived from librispeech for text-to-speech.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1097–1100.

A   Training Data

To train SpeechSim, we use LibriLight (Kahn et al., 2020), an unlabelled dataset containing approximately $60,000$ hours of self-recorded audiobooks. To train projection models on top of SpeechSim, we use different datasets depending on the target facet. Specifically, the projection models related to speaker voices and gender are trained on LibriTTS train-other (Zen et al., 2019). The projector related to emotion is trained on EmoV (Adigwe et al., 2018) and Expresso (Nguyen et al., 2023) train splits. For Expresso, we selected audios with a single speaker and removed narration and enunciated classes. Expresso and EmoV datasets have only the neutral class in common.

Next, the accent projection model is trained on VCTK-train (Veaux et al., 2016), which contains 12 English accents spoken by 110 speakers. Finally, the background noise projection model is trained on AudioSet (Gemmeke et al., 2017). Here, we select all examples that are tagged as "Speech" and which have between 1 and 3 additional tags as we found examples with more tags to be too noisy. We randomly sample 70% of the selected examples to build the train set, the remaining 30% are randomly split as dev and test.

B   Baseline Representations

B.1   Generic representations

Wav2Vec-BERT   We experiment with representations provided by a 600M-parameter Wav2Vec-BERT model (Chung et al., 2021a). We take embeddings of intermediate layers, with each utterance represented as a sequence of 1024-dimensional vectors corresponding to 40ms frame.

HuBERT   We use HuBERT embeddings, produced by the Base model (Hsu et al., 2021) (95M parameters), trained on LibriSpeech (Panayotov et al., 2015). We follow Lakhotia et al. (2021) and use embeddings extracted from the 6th layer. This model has a framerate of 20 ms.

SoundStream Another type of representations found to be useful in AudioLM is Sound-Stream (Zeghidour et al., 2021). While Wav2Vec-BERT representations were found to be high-level, SoundStream representations are fine-grained, faithfully representing minute details of speech (Borsos et al., 2023a). We use a version of SoundStream with 60 residual quantization layers, each quantized into 1024 tokens and 200 ms frames. The encoder has around 12M parameters. For uniformity, we use dequantized representations. This model operates on 24kHz audio.

B.2 Specialized representations

Training from scratch To evaluate the contribution of the pretrained model SpeechSim, we also trained a model from scratch for each facet on the labelled data directly. These models have the same architecture as SpeechSim but with 6 layers. They are trained with the same contrastive objective used for training the projection models.

C Effect of decoding method and temperature on acoustic diversity

C.1 SPEAR-TTS: Effect of Best-of-K decoding

Kharitonov et al. (2023) leveraged the stochasticity of the temperature decoding to improve audio quality of the generated speech in an AudioLM-like model. Namely, they proposed to sample K examples for the same input and then select the sample with the highest acoustic quality as the output. Now we check if that approach brings a penalty to the acoustic diversity.

Since SPEAR-TTS is only stochastic in the semantic-to-acoustic token decoding, we use the same task and AudioLM model as above. Again, for each semantic token sequence, we sample 128 acoustic token sequences that correspond to it. Next, we randomly split them in groups of $K \in \{1, 2, 4\}$ and, within each group, select one with the highest acoustic quality score, relying on the same model as Kharitonov et al. (2023). After this filtering, each semantic token sequence is mapped into 128, 64, or 32 audio samples (for K=1, 2, and 4). In the first two cases, we sample 32 sequences so that all sets have the same size.

From Table 9 see that, except for the gender facet, higher K always brings a diversity penalty.

K	Voice	Gender	Emotion	Accent	Background Noise
1	64	41	62	63	61
4	0	27	0	0	0

Table 9: Best-of-K filtering on the semantic-to-acoustic tokens resynthesis task. The proportion (%) of utterances where Best-K with $K \in \{1, 4\}$ have higher diversity scores than with $K = 2$.

T	Voice	Gender	Emotion	Accent	Background Noise
0.7	12	33	21	18	29
0.9	45	21	40	40	36

Table 10: Count (out of 64) of instances where sets of utterances generated with $T \in \{0.7, 0.9\}$ have higher diversity scores than with $T = 0.8$.

C.2 Text Synthesis: Influence of Temperature

Temperature is a parameter aiming at controlling the sampling distribution. As the sampling temperature approaches zero, samples for a fixed input become increasingly deterministic; while a high temperature introduces diversity in the sampling process. A priori, one can expect that this token-level diversity can translate to higher-level, per-utterance acoustic diversity. In this experiment we use a single-stage fully auto-regressive TTS system that maps grapheme text representations to a sequence of acoustic tokens, akin to VALL-E (Wang et al., 2023). This model is trained on a combination of Multilingual LibriSpeech (Pratap et al., 2020), LibriLight (Kahn et al., 2020) and other public data. Table 10 shows that voice diversity is by far the most affected by small variations of temperature while gender and background noise are barely impacted. Samples with low temperature are also less diverse in terms of emotion and accent.

D Additional experiments on testing whether facets are mixed up together

D.1 Emotions and Voices

We examine the interactions between emotion and voice diversity within metrics as explained in Section 7.2. The number of speakers is 1 or 4 depending on the diversity level of the set of utterances. For sets with high entropy (i.e. high diversity), there is an unique speaker voice while there are 4 speakers for sets with low entropy (i.e. low diversity). We report the results in Table 11. We observe that SpeechSim/Emotion is unaffected by the changes in the number of speakers, while Speech-Sim/Voice follows the speaker diversity and hence

| | EmoV | |
	Avg. cosine	Vendi score
SpeechSim	-0.384 (± 0.347)	-0.372 (± 0.347)
SpeechSim/Voice	-0.609 (± 0.185)	-0.451 (± 0.287)
SpeechSim/Emotion	0.970 (± 0.042)	0.985 (± 0.028)
	Expresso	
	Avg. cosine	Vendi score
SpeechSim	-0.257 (± 0.423)	-0.218 (± 0.423)
SpeechSim/Voice	-0.397 (± 0.464)	-0.290 (± 0.530)
SpeechSim/Emotion	0.994 (± 0.013)	0.985 (± 0.019)

Table 11: Changing emotion and speaker diversity in the opposite directions. Average Spearman correlation between emotion diversity and diversity scores induced by speech representations.

	Avg. cosine	Vendi score
SpeechSim	-0.677 (± 0.221)	-0.737 (± 0.119)
SpeechSim/Voice	-0.743 (± 0.199)	-0.790 (± 0.110)
SpeechSim/Accent	0.795 (± 0.159)	0.999 (± 0.008)

Table 12: Changing accent and voice diversity in the opposite directions. Average Spearman correlation between accent diversity and diversity scores induced by speech representations.

it has a negative correlation with the emotion diversity score. SpeechSim focuses on the changes in the speaker diversity, thus having a negative correlation with the gender diversity.

D.2 Varying accents and voices in opposite directions

In this experiment, we use 7 voices for sets with high entropy and 30 voices for sets with low entropy. The results are reported in Table 12. We see that the projected metric is unaffected by changes in voices, while the vanilla version of SpeechSim is focused on the speaker facet.

E Acoustic diversity in Text-to-Speech systems - Model details

Bark TTS[3] is a generative text-to-audio system that is based on a decoder-only Transformer model that operates on discretized speech representations, akin to AudioLM (Borsos et al., 2023a) and VALL-E (Wang et al., 2023). Bark TTS synthesizes utterances in a random voice and acoustic conditions.

StyleTTS 2 (Li et al., 2024) combines a style diffusion model with speech with WavLM-based adversarial training (Chen et al., 2022a) for an end-to-end training of a multi-component system. We use a checkpoint trained on LibriTTS (Zen et al., 2019). The samples are conditioned on a female voice, but randomize timbre and prosody.[4]

Tortoise TTS (Betker, 2023a) combines autoregressive and diffusion-based components (Betker, 2023b). For training, it pooled LibriTTS (Zen et al., 2019), HiFiTTS (Bakhturina et al., 2021), and public data sourced from audiobooks and podcasts.

FastSpeech 2 (Ren et al., 2020) consists of a Transformer-based encoder that embeds phoneme-based text representation, a variance adapter that predicts pitch, energy, and duration values, and a non-autoregressive decoder. We use a publicly available re-implementation trained on LJSpeech (Ito and Johnson, 2017).[5]

---

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

[4] https://colab.../Inference_LibriTTS.ipynb
[5] https://github.com/ming024/FastSpeech2