# MODELING ABSTRACT STYLE PROMPTS FOR TEXT-TO-SPEECH MODELS

Anonymous authors

004

010 011

012

013

014

015

016

017

018

019

021

023

025

026

027 028 029

030

Paper under double-blind review

#### ABSTRACT

A recent trend in text-to-speech synthesis (TTS) is to construct models capable of generating naturalistic speech that adheres to a textual style prompt describing the speaker's voice and speaking style. In this paper, we propose a crisper definition of style-prompted TTS by categorizing style tags by how they can be collected (automatic tags obtainable using signal processing tools e.g. low-pitched and slow; *demographic* tags obtainable using speaker demographics e.g. male and American accent; and *abstract* tags which need human-annotations e.g. authoritative and awed) and what they represent (*intrinsic* tags inherent to speaker identity e.g. gender, average pitch, texture; and *situational* tags specific to utterance-level speaking styles e.g. emotion). Compared to previous work, we expand the space of style prompts substantially by covering 47 abstract tags, 10 demographic tags and 6 automatic tags. For abstract intrinsic tags, we annotate a subset of speakers from the VoxCeleb (Nagrani et al., 2020) dataset. For abstract situational tags, we leverage existing speaking-style-based datasets Expresso (Nguyen et al., 2023) and EARS (Richter et al., 2024). We train a style-prompted TTS model based on Parler-TTS (Lyth & King, 2024; Lacombe et al., 2024b) using these datasets and find that our model outperforms baselines on speech-style consistency metrics. Our collected dataset and model will be open-sourced.

#### 1 INTRODUCTION

Text-to-speech systems that are controllable by natural language text style prompts a.k.a. style-prompted TTS systems e.g. (Guo et al., 2022; Ji et al., 2024; Leng et al., 2023; Vyas et al., 2023; Lacombe et al., 2024b; Jin et al., 2024) have been gaining prominence in the past few years. Rather than providing control via a few seconds of reference speech (Peng et al., 2024; Wang et al., 2023) exhibiting the desired style, these models allow users to do so via natural language instead, which provides a more explicit, intuitive, and privacy-preserving control medium. Training these models
requires a dataset which has speech utterances annotated with natural language style prompts.

038 When humans describe speech speaking styles in natural language, they do so with a rich and diverse vocabulary spanning a wide range of style tags covering aspects like pitch, texture, emotion 040 and rhythm and more. We propose a crisper definition of style-prompted TTS that rigorously catego-041 rizes style tags along two axes: the mechanism by which one can obtain tag annotations (automatic, 042 demographic and abstract tags) and what speech aspects the tag represents (intrinsic tags and situa-043 tional tags). Based on an extensive survey of previous style-prompted TTS work (Section 2), we find 044 that existing work offers natural-language control over some, but not all of these categories, often overlooking the importance of covering more abstract speech style tags that cannot be automatically extracted. More importantly, the only current open-source model, Parler-TTS (Lyth & King, 2024; 046 Lacombe et al., 2024b) only supports automatic tags, which motivates the need for an open-source 047 model that can support all categories. 048

We create a list of 63 style tags consisting of 26 abstract intrinsic, 21 abstract situational, 10 demographic and 6 automatic tags covering a wide variety of speech styles. As described in Section 3, to support all these tags, we collect abstract intrinsic human annotations for a subset of the Vox-Celeb (Nagrani et al., 2020) dataset, creating StyledVoxCeleb. While not originally proposed for style-prompted TTS, Expresso (Nguyen et al., 2023) and EARS (Richter et al., 2024) cover a rich variety of abstract situational speaking styles and hence we reuse them for TTS. We finetune Parler-



Figure 1: An overview of our data collection procedure.

TTS (Lacombe et al., 2024b; Lyth & King, 2024) on these datasets and find (Section 5) that our model outperforms competitive baselines on speech-style consistency metrics.

In summary, our contributions are:

- We provide a crisper categorization of style tags and perform an extensive survey of prior work based on this categorization.
- We create StyledVoxCeleb, a subset of the VoxCeleb (Nagrani et al., 2020) dataset annotated with abstract intrinsic style tags, reuse Expresso and EARS for TTS, and train style-prompted TTS model that cover all categories of speech style tags.
- We demonstrate that training on our dataset pool results in improved performance on speech-style consistency metrics, obtaining +0.1 in consistency MOS and +0.06 tag recall (metrics introduced in Section 4) as compared to the next best baseline.

We will open-source our model and collected data upon publication.

084 085

087

068

069

073

074 075

076 077

079

081

082

083

#### 2 BACKGROUND AND MOTIVATION

We can describe speech styles with a rich and diverse vocabulary, capturing aspects such as pitch, emotion, rhythm, speaking rate and more. We draw a distinction between *intrinsic* tags that are tied 090 to a speaker's identity and persist across all utterances belonging to that speaker (e.g. average pitch 091 and vocal texture) and *situational* tags that describe the speaking style (e.g. emotion) of individual 092 utterances.<sup>1</sup> This distinction is important for deciding a data collection strategy; while intrinsic tags can be annotated on a per-speaker basis, situational tags must be annotated on a per-utterance basis, 093 which is significantly more expensive. Furthermore, depending on how the tags can be collected, 094 we distinguish between *abstract* tags that are complex and require human annotations (e.g. clarity, 095 texture, emotion), demographic tags obtainable from speaker demographics (e.g. gender and accent) 096 and *automatic* tags  $^2$  that are obtainable via signal processing tools (e.g. pitch/F0, energy, speaking 097 rate). Abstract and automatic tags can be both intrinsic and situational, while demographic tags are 098 always intrinsic since they pertain to speaker demographics. In the rest of this paper, we classify tags into four categories: abstract intrinsic, abstract situational, demographic and automatic tags 100 (combining intrinsic and situational automatic tags).

101 102 Style-prompted TTS should support a diverse space of speech style prompts, covering many tags 103 across all categories. We perform a comparison of prior work summarized in Table 1 and make 104 several observations. First, we notice that none of the previous datasets apart from AudioBox and

105

<sup>2</sup>While there exist automatic emotion classifiers (Ma et al., 2023) for a subset of emotions used by prior work e.g. (Jin et al., 2024), we found that their quality on our datasets is unsatisfactory. In this work, we still consider them to be abstract tags.

<sup>&</sup>lt;sup>1</sup>Some tag categories can be both intrinsic and situational and need to be handled carefully; see Appendix A.

Dataset		Abst Dem		Auto	# Abst	<b>Open-Sourced</b>			
	Intr	Sit	-			Data	Model		
PromptSpeech (Guo et al., 2022)	×	1	1	1	4	1	×		
NLSpeech (Yang et al., 2023)	X	1	1	1	?	X	X		
PromptStyle (Liu et al., 2023)	X	1	1	1	?	X	X		
TextrolSpeech (Ji et al., 2024)	X	1	1	1	8	1	X		
Coco-Nut (Watanabe et al., 2023)	1	1	1	1	?	1	X		
PromptTTS2 (Leng et al., 2023)	X	X	1	1	0	X	X		
MEAD-TTS (Guan et al., 2024)	X	1	1	1	8	1	X		
AudioBox (Vyas et al., 2023)	1	1	1	1	?	X	X		
ParlerTTS (Lacombe et al., 2024b)	X	X	1	1	0	1	1		
LibriTTS-P (Kawamura et al., 2024)	1	X	1	1	44	1	×		
SpeechCraft (Jin et al., 2024)	X	1	1	1	7	1	X		
Ours	1	1	1	1	47	$\checkmark$	1		

Table 1: A comparison of existing style-prompted TTS papers. The # Abst column denotes the number of abstract tags in each dataset. We denote with the ? symbol those datasets whose abstract tag count is unknown. To the best of our knowledge, only AudioBox and Coco-Nut cover all three abstract tag categories. Of these, the AudioBox dataset is closed-source and Coco-Nut is only 8 hours long, making both unusable for training a TTS model. While LibriTTS-P (Kawamura et al., 2024) has nearly as many tags as ours, it does not cover abstract situational tags.

129 130

131 Coco-Nut cover all categories. Both are unusable for training TTS models, since AudioBox is closed-source and Coco-Nut is only 8 hours long. This motivates the need for a new, open-sourced 132 TTS model that can take handle tags from all categories, especially since the only open-source 133 model, Parler-TTS, does not support any abstract tags at all. Second, all of the datasets that support 134 abstract situational tags (emotions) only cover a maximum of 8 tags, motivating the need to sub-135 stantially expand the space of abstract situational tags. Thirdly, while some of these datasets (e.g. 136 TextrolSpeech) start with a limited set of tags (e.g. automatically extracted pitch, speaking rate and 137 volume) and synthetically expand their style prompt vocabulary (e.g. by using LLMs to rewrite style 138 tags with synonyms) to better mimic how humans would describe speech styles, this does not add a 139 real signal to the dataset. 140

We resolve these issues by emphasizing our focus on a variety of abstract style tags that are difficult to extract using automatic extractors, requiring human annotations. We manually create a list of 47 abstract tags (26 intrinsic, 21 situational) covering pitch, texture, clarity, volume and rhythm for intrinsic and emotion and expressiveness for situational tags. Combined with 10 demographic tags (2 gender, 8 accents) and 6 automatic tags (3 pitch levels and 3 speaking rate levels), we cover a total of 63 tags; the full list is available in Appendix A. We set up 4 datasets aiming to target as many of these 63 tags as possible, and describe their creation in Section 3.

147 148

#### 3 DATASETS

3.1 DATA COLLECTION

153 We present an overview of our dataset collection procedure for each tag category in Figure 1. We 154 set up four datasets: (a) StyledVoxCeleb, a subset of the VoxCeleb (Nagrani et al., 2020) dataset 155 we annotate with abstract intrinsic tags, (b) Expresso (Nguyen et al., 2023) and EARS (Richter 156 et al., 2024), two existing expressive speech datasets whose speaking styles we remap to abstract 157 situational tags, and (c) a 150-hr subset of LibriTTS-R (Koizumi et al., 2023) we call LTTSR-150 annotated with demographic and automatic tags. We annotate StyledVoxCeleb, Expresso and EARS 158 with demographic and automatic tags as well. Preprocessing information for each dataset can be 159 found in Appendix C. Across all datasets, every audio clip's style tag annotations are converted to 160 a natural language style prompt using a Mistral (Jiang et al., 2023) LLM prompted with a comma-161 separated list of style tags and instructed to generate a style prompt (details in Appendix D). Every

Dataset	# Spkr	# Utts.	Dur.
StyledVoxCeleb	596	116k	256.08h
Expresso	4	16k	30.21h
EARS	107	15k	60.58h
LTTSR-150	2410	95k	178.52h

165		Expresso	4	16k	30.21h	
166		EARS	107	15k 051-	60.58h	
167		LI15K-150	2410	93K	178.320	
168		Tabla	2. Detect	atatistica		
169		Table	2. Dataset	statistics.		
170						
171	example in our datasets	thus consists of (a	) an audio c	lin (h) a	text style prom	nt generated from th
172	annotated style tags and	(c) a text transcrit	tion.	.np, (0) u	text style pron	ipt generated from a
173	uniforme a style tags and					
174	Abstract Intrinsi	C TAGS We crea	te the Style	edVoxCel	eb dataset by a	annotating a subset
175	VoxCeleb (Nagrani et a	al., 2020) (a datase	t consisting	of natura	l. in-the-wild	speech from YouTul
176	celebrity interviews wit	h high speaker dive	ersity spanni	ing accen	ts, ages and eth	inicities) with abstra
177	intrinsic tags by hiring	workers on Amazo	n Mechanic	al Turk.	We apply this a	annotation to all utte
178	ances spoken by that sp	eaker. This data col	llection effo	rt is comp	plementary to p	rior work (Kawamu
179	et al., 2024) that collect	ted such data for th	e LibriTTS-	-R (Koizu	imi et al., 2023	) dataset. We show
180	Section 5 that our datas	et outperforms Lib	riTTS-P wh	ien evalua	ated for speech	-style consistency.
181						
182	Quality Control We	provide a qualifica	tion task to	Amazon	Mechanical T	urk workers to chec
183	their ability to understa	nd style tags. The	task consist	s of 6 ma	nually selected	I pairs of speech clij
184	where one exhibits a sty	approximation one doesn't	. we ask an	notators	to select which	one exhibits the sty
185	and keep only mose 38		ceeded off a	at least 5 (	examples, deta	iis iii Appendix E.
186	Collecting Annotation	. Civen e encelte	#		mtativa audia	fla consisting of my
187	tiple utterances (3 – 8	cline whose total	I, we create	20 - 40	seconds) concr	tensted together W
188	provide this audio file	the speaker's name	a and a list	of our in	trinsic speech	style tags with defin
189	tions (see Appendix A)	to annotators on A	mazon Me	chanical '	Furk and ask th	to write at least
190	distinct style tags. Our	annotation UI can	be viewed a	at Append	lix E. For ever	y celebrity, we colle
191	5 annotations. We obse	erve that the annot	ations are v	ery subje	ctive and diffe	rent annotators sele
192	different tags for the sa	me celebrity. The	refore, we k	keep only	those tags that	t at least 2 annotato
193	agree on in our train and	d dev set, and only	those that a	it least 3 a	annotators agre	e on in our test set.
194						
195	Selecting Celebrities	We expect famou	s or distinct	tive celeb	rities to be mo	re familiar to annot
196	tors. We select such a su	bset using three lo	ose heuristic	cs: (a) we	parse an IMD	list of 163 celebriti
197	with distinctive voices	and find 39 in Voz	xCeleb, (b)	we ask C	hatGPT to nam	e 300 celebrities wi
198	distinctive voices and fi	nd 112 in VoxCele	$\mathbf{x}$ , and $(\mathbf{c})$ we	e find Wil	kipedia pages f	or VoxCeleb celebri
199	names assuming page 1	ength is a prove for	or fame Co	200 cele	all three source	es and accounting f
200	overlap we obtain a list	t of 302 celebrities	After colle	cting ann	otations for the	es celebrities we fir
201	that the style tag distrib	ution is imbalance	d. with 12 ta	ags <sup>5</sup> hav	ing fewer than	5000 annotated clin
202	We use GPT-4 (OpenA)	I et al., 2024) to ob	tain a rough	n list of ce	elebrities that a	re likely to have the
203	tags by instructing it to	output a list of sty	le tags that a	are assoc	ated with a cel	ebrity's voice (detai

162 163 164

214

1 a-1 es 1 th 1 ty lia 1 or 2 nd 2 os. 2 se 2 ils 204 in Appendix D) for every celebrity in VoxCeleb. Since we can only provide the celebrity's name 205 rather than the actual speech clip, GPT-4 needs to rely on its parametric knowledge base in order 206 to complete this task; while imperfect, it may still provide some signal towards which celebrities 207 to target. We select a maximum of 40 celebrities per tag, ending up with a list of 187 additional 208 celebrities to annotate (most tags have far fewer than 40 celebrities labelled by GPT-4). Finally, we annotate 107 additional celebrities, resulting in a total of 596 celebrities for StyledVoxCeleb. 209

210 We split every speaker in StyledVoxCeleb into train (80%), dev (10%), and test (10%), ensuring that 211 there is no transcript overlap across splits. 212

<sup>213</sup> <sup>3</sup>https://www.imdb.com/list/ls001839542/

<sup>&</sup>lt;sup>4</sup>https://github.com/martin-majlis/Wikipedia-API

<sup>&</sup>lt;sup>5</sup>lisp, hushed, pitchy, staccato, monotonous, punctuated, vocal fry, guttural, singsong, soft, stammering, 215 shrill

216 **ABSTRACT SITUATIONAL TAGS** We reuse Expresso (Nguyen et al., 2023) and EARS (Richter 217 et al., 2024), two existing expressive speech datasets that consist of speakers acting out various 218 emotions and speaking styles. Expresso contains 4 speakers while EARS contains 107. We filter 219 out neutral and non-speaking utterances in both datasets, lightly preprocess them (details in Ap-220 pendix C) and then label each utterance by simply mapping the reading styles in each dataset to our tag vocabulary (details in Table 6). For example, the projected style in Expresso gets mapped to the 221 tag loud. We split the Expresso dataset into train (80%), dev (10%), and test (10%), ensuring that 222 there is no transcript overlap across splits. Some utterances in EARS do not have emotion labels; 223 we place them all into the train set. We split the utterances that have emotion labels into train (80%), 224 dev (10%), and test (10%), ensuring overall that there is no transcript overlap across splits. 225

DEMOGRAPHIC AND AUTOMATIC TAGS We use either dataset metadata or GPT-4 for obtaining accent and gender, and automatic signal processing tools for extracting pitch and speaking rate. Following previous work (Lyth & King, 2024), we also extract the noise level of the audio (this is not a *style tag*; rather, it aids the model to differentiate between clean and noisy speech). We extract these tags for all 3 datasets: StyledVoxCeleb, Expresso and EARS. Additionally, we annotate a 150-hr subset of the train split of the LibriTTS-R (Koizumi et al., 2023) dataset (along with its dev and test sets) we call LTTSR-150 with gender, pitch, speaking rate and noise levels.

**Gender and Accent** For StyledVoxCeleb, we prompt GPT-4 with the name of the celebrity and ask it to output the celebrity's gender and accent. <sup>6</sup> The prompt and generation details are available in Appendix D. We use dataset metadata for Expresso, EARS and LibriTTS-R (for EARS, we use the 'native language' column of the dataset as a proxy for accent).

Pitch, Speaking Rate and Noise Levels We use the Dataspeech (Lacombe et al., 2024a) library 239 to label our datasets with pitch, speaking rate, and noise levels. For pitch, we use PENN<sup>7</sup> using 240 default hyperparameters and compute the mean pitch across all utterances of a given speaker. Then, 241 we apply gender-dependent thresholds to label each speaker with low-pitched (male: < 115.7 Hz, 242 female: < 141.6 Hz), high-pitched (male: > 149.7 Hz, female > 184.5 Hz) or medium-pitched 243 (male: 115.7 Hz < x < 149.7 Hz, female: 141.6 Hz < x < 184.5 Hz); these thresholds are 244 gender-dependent since humans perceive male speakers to have lower pitch than female speakers on 245 average. For speaking rate, we use  $g_{2p}^{8}$  to convert the text transcription to phoneme transcriptions 246 and then use the number of phonemes per second (PPS) as the speaking rate. We apply thresholds to 247 label each utterance with slow (< 11.5 PPS), fast (> 19.1 PPS) and measured (11.5 PPS < x < 19.1248 PPS). Finally, for noise levels, we use the signal-to-noise ratio (SNR) extracted using Brouhaha<sup>9</sup> and use Parler-TTS (Lacombe et al., 2024b)'s noise bins to assign each utterance one of the following 249 noise tags: very noisy, quite noisy, slightly noisy, moderate ambient sound, slightly clear, quite clear, 250 very clear. 251

251 252 253

254

255

256

257

233

238

#### 3.2 DATASET STATISTICS

We report dataset statistics for each dataset we setup (StyledVoxCeleb, Expresso, EARS and the LibriTTS-R subset LTTSR-150) combining train, dev and test splits in Table 2. We report the distribution of each accent tag in Figure 2 and abstract tags in Figure 3. The distribution of gender tags is 56.6% male, 43.4% female, of pitch tags is 21.4% low-pitched, 41.5% medium-pitched and 37.05% high-pitched, and of speaking rate tags is 12.6% slow, 75.7% measured and 11.6% fast.

262

263

264

265

266

267

268

269

## 4 EXPERIMENTAL SETUP

We use the Parler-TTS (Lyth & King, 2024; Lacombe et al., 2024b) model as our backbone for all experiments. Parler-TTS is a style-prompted TTS model trained on 45K hours of data, consisting of the English split of Multilingual Librispeech (Pratap et al., 2020) and LibriTTS-R (Koizumi et al., 2023) annotated with automatic style tags.

<sup>&</sup>lt;sup>6</sup>We manually verified a subset of the generated metadata and found it to be of high quality.

<sup>&</sup>lt;sup>7</sup>https://github.com/interactiveaudiolab/penn

<sup>&</sup>lt;sup>8</sup>https://github.com/roedoejet/g2p

<sup>&</sup>lt;sup>9</sup>https://github.com/marianne-m/brouhaha-vad

Model Architecture The crux of Parler-TTS is an autoregressive decoder speech language model
that generates DAC (Kumar et al., 2023) audio tokens. To condition on text transcripts, the text
transcript is tokenized using the Flan-T5 (Chung et al., 2022) tokenizer, passed through a linear
embedding layer, and prepended to the input sequence of the decoder. To condition on the text
style prompt, the text encoder, a frozen Flan-T5 model, maps the text style prompt to a sequence of
hidden-state representations that are attended to via cross-attention layers in the decoder.

Training We initialize our model with the parler-tts/parler-tts-mini-v1 open-source checkpoint and use the official Parler-TTS <sup>10</sup> library to finetune on the training splits of the 4 datasets we set up; StyledVoxCeleb, Expresso, EARS and LTTSR-150. We train on 4 NVIDIA A40 GPUs with a batch size of 8 and 2 gradient accumulation steps. We train for 9 epochs with a non-warmup cosine learning rate scheduler, a peak learning rate of 0.00008 and a weight decay of 0.01.

282 283

276

Inference We perform inference runs using the default Parler-TTS generation hyperparameters (temperature 1.0, repetition penalty 1.0, 2580 total tokens). Since autoregressive TTS is prone to decoding instabilities, we attempt to mitigate this by retrying inference a maximum of 3 times, stopping when the WER between the ASR transcript of the generated sample and the input text (using the same setup as our WER evaluation metric) falls below 20 or choosing the sample with the lowest WER out of the 3 generated samples.

289 290

291

305

4.1 EVALUATION DATASET

We start by combining the test splits of StyledVoxCeleb, Expresso, EARS and LibriTTS-R. For each tag in our tag vocabulary, we find a maximum of 5 clips that have been annotated with that tag and select them for inclusion in our evaluation dataset. For each clip, we refer to this tag as its *tag of interest*. We randomly select pitch and speaking rate with a 50% probability for inclusion along with the tag of interest, gender, noise level and then generate a style prompt from these tags for use in evaluation. Our final evaluation dataset consists of 298 clips.

# 298299 4.2 EVALUATION METRICS

We use metrics that evaluate the speech clip for three desiderata: speech quality, content correctness and speech-style consistency. For human evaluation metrics, we use annotators recruited on Amazon Mechanical Turk; Appendix E contains details about our annotation user interfaces and annotation costs. For every human evaluation metric, we collect 3 human annotation scores per test dataset item. We report the mean and 95% confidence intervals of the MOS scores (Ribeiro et al., 2011).

Speech Quality Following previous work (Vyas et al., 2023; Kawamura et al., 2024), we compute
 a Naturalness MOS metric where each human annotator is provided speech clips and asked to rate
 its naturalness and realisticity (human-likeness) on 5-point Likert scales.

 Content Correctness We report a WER metric that computes the Word Error Rate (WER) between (a) the ASR transcript of the speech clip and (b) the input transcript, after applying a text normalizer to both texts. We use the distil-whisper/distil-large-v2 (Gandhi et al., 2023) model for ASR, and Whisper<sup>11</sup> for text normalization.

Speech-Style Consistency Following Kawamura et al. (2024) and Ji et al. (2024), we report a Consistency MOS metric where each human annotator is provided a speech clip and the input style prompt and asked to rate the consistency between the two on a 5-point Likert scale.

In addition, we report fine-grained tag-level evaluation. Instead of evaluating adherence to the whole
style prompt (e.g., *A woman's speech is delivered slowly with a high-pitched tone, expressing dis- gusted emotions, in a clear and quiet environment*), we ask annotator to select style tags they hear.
For example, for the same example, annotator might select *female, disgusted* as pronounced style.

<sup>322</sup> 323

<sup>&</sup>lt;sup>10</sup>https://github.com/huggingface/parler-tts

<sup>&</sup>lt;sup>11</sup>https://github.com/huggingface/transformers/blob/main/src/ transformers/models/whisper/english\_normalizer.py

For each tag, we compute its recall (fraction of instances in which the relevant tag was selected by the annotator), and report the average tag recall as well as per-category average tag recall.

For two automatic tag types (pitch and speaking rate), we further report an Accuracy score. We run the generated speech clip through the same pitch and speaking rate extractors used to build our datasets to obtain predicted style tags. We use the style prompt's gender to decide which pitch bins to use and label each generated utterance individually, rather than speaker-level mean aggregation used for building the datasets. We compute the speaking rate from the phoneme sequence obtained from the ASR transcript of the generated speech. We then compare the predicted tags pitch with the desired tags in the input style prompt, giving a score of 1 if the labels match and 0 otherwise.

4.3 BASELINES

Due to the absence of open-source style-prompted TTS models other than Parler-TTS, all our baselines finetune Parler-TTS on different datasets with the same training and inference setup as ours.

338 339 340

345

346

347

348

349

350

351

352 353

354

334

335

**Init.** This is the Parler-TTS model that we initialize all models with.

+LTTSR We finetune Parler-TTS on the LibriTTS-R (Koizumi et al., 2023) dataset. We extract gender tags using dataset metadata and automatic tags using our signal processing pipeline for extracting pitch, speaking rate and noise levels. While Parler-TTS is already trained on LibriTTS-R, it uses different binning thresholds for pitch and speaking rate; this baseline ablates that mismatch.

+LTTSP,Exp,EARS We finetune Parler-TTS on a combination of existing datasets that cover all tag categories: Expresso and EARS for abstract situational tags and LibriTTS-P (Kawamura et al., 2024) which annotates the LibriTTS-R dataset with abstract intrinsic tags. LibriTTS-P provides 3 annotations (each consisting of a list of style tags) per speaker and each style tag optionally has one of two qualifiers (*slightly* and *very*) that indicates the strength of the style tag. We preprocess the annotations by removing tags with the *slightly* qualifier and remapping some style tags to those in our vocabulary (see Appendix C). For each clip in the dataset, we select one of the three annotations corresponding to its speaker at random and combine with automatic tags extracted using our pipeline. This baseline ablates the use of LibriTTS-P versus our StyledVoxCeleb dataset.

#### 5 RESULTS

Model	Cons. MOS $\uparrow$		Ta	ig Recal	l <b>l</b> ↑		Accura	acy % ↑
		All	Intr	Sit	Dem	Auto	Pitch	Rate
GT	$3.76\pm0.57$	0.62	0.56	0.62	0.74	0.70	56.52	93.24
Init.	$3.14\pm0.44$	0.25	0.23	0.16	0.28	0.59	62.73	77.01
+LTTSR	$3.15\pm0.47$	0.26	0.20	0.19	0.33	0.58	73.91	66.21
+LTTSP,Exp,EARS	$3.19\pm0.31$	0.30	0.24	0.29	0.26	0.60	72.67	75.00
Ours	$3.29 \pm 0.40$	0.36	0.29	0.29	0.52	0.61	72.05	75.68

Table 3: Speech-Style Consistency results comparing various baseline models and ours. We report the mean and 95% confidence intervals for Consistency MOS. Tag recalls are averaged across all tags (All) and broken down by each tag category (Intr. is abstract intrinsic, Sit. is abstract situational, Demo. is demographic and Auto. is automatic). We find that our model outperforms baselines at consistency MOS and overall Tag Recall.

370 371 372

**Speech-Style Consistency** Table 3 reports model performance along various metrics that aim to evaluate how well the generated speech adheres to the provided text style prompt. The consistency MOS ranges from 1 - 5, the tag recall from 0 - 1 and the accuracies from 0 - 100%. Our model achieves the highest consistency MOS score, verified by running a paired bootstrap significance test comparing the two highest MOS scores (ours and the +LTTSP,Exp,EARS baseline) that finds the difference is statistically significant with a p-value of 0.004. Furthermore, the Tag Recall

355 356

368

378	Model	NMOS $\uparrow$	WER↓
380	GT	$3.94\pm0.42$	8.10
381	Init.	$3.05\pm0.25$	4.91
382	+LTTSR	$3.12 \pm 0.20$	4.75
383	+LTTSP,Exp,EARS	$2.99\pm0.18$	6.33
384	Ours	$2.80 \pm 0.16$	9.12
00E			

Table 4: Speech Quality and Content Correctness results. We report the mean and 95% confidence intervals for Naturalness MOS.

386

387

390 scores provide a more finegrained understanding of model performance. We outperform all base-391 lines, including the LTTSP, Exp, EARS baseline on intrinsic tags, showing the benefits of training on StyledVoxCeleb versus LibriTTS-P. Since both our model and the LTTSP,Exp,EARS baseline 392 is trained on Expresso and EARS, we match performance on situational tags, but outperform other 393 baselines, demonstrating the benefits of training with Expresso and EARS. Additionally, our model 394 outperforms on demographic tags as well due to the presence of a rich diversity of accents in Styled-395 VoxCeleb. When automatic tags (pitch and speaking rate) are evaluated via automatic accuracy 396 scores, we find that the baselines trained without any abstract tags (Init. and +LTTSR) slightly out-397 perform our model by about 2%. However, all models perform similarly at automatic tags when 398 evaluated using tag recall, showing that humans do not exhibit strong preferences between models 399 when evaluating pitch or speaking rate. We note that the ground truth pitch accuracy is unusually 400 low because of a mismatch between how pitch is computed during evaluation versus dataset con-401 struction: the pitch is computed on an utterance-level basis during evaluation, while it was obtained on a speaker-level basis when constructing the style prompt in the dataset. 402

403

404 **Speech Quality and Content Correctness** Table 4 compares the naturalness and content correct-405 ness of the generated speech across models. We find that the models trained without any abstract tags 406 (+LTTSR and Init.) widely outperform our model and the +LTTSP,Exp,EARS baseline on both nat-407 uralness and WER. Training on LibriTTS-P, Expresso and EARS, despite being clean, high-quality audio data worsens both WER and naturalness. We hypothesize this is due to the introduction of 408 speaking styles that are harder to transcribe and the relatively small scale of the Expresso and EARS 409 data. Furthermore, training on StyledVoxCeleb (Ours) worsens it further, which we hypothesize is 410 due to the presence of noisier in-the-wild speech (VoxCeleb) in our training data, which introduces 411 speech artifacts in the generated speech. This is a limitation of the audio quality and size of our 412 dataset, which we expect will be mitigated as we scale our dataset to more speakers (for example, 413 Voicecraft (Peng et al., 2024), a voice cloning TTS model trains on large-scale, in-the-wild noisy 414 data and achieves low Word Error Rates). The WER of Init. and LTTSR is substantially lower than 415 the ground truth; this is likely because both models are trained on read audiobook data which is 416 easier for humans and ASR systems to understand.

417 418

## 6 RELATED WORK

419 420

Style-Prompted Text-to-Speech Models Table 1 already compares several existing style-421 prompted text-to-speech papers with respect to our tag categorizations. PromptTTS (Guo et al., 422 2022), one of the first papers to introduce style-prompted TTS, consists of 4 emotions and automatic 423 tags and is trained on a synthetic emotion dataset; PromptTTS2 (Leng et al., 2023), a successor fo-424 cuses on an improved model architecture. Other emotion-focused work includes InstructTTS (Yang 425 et al., 2023), PromptStyle (Liu et al., 2023) and MEAD-TTS (Guan et al., 2024) which focus on 426 collecting or annotating emotional data using human voice actors or annotators. TextrolSpeech (Ji 427 et al., 2024) also focuses on emotion by collating several existing emotion classification datasets 428 for use for style-prompted TTS. Recently, Parler-TTS (Lacombe et al., 2024b; Lyth & King, 2024), 429 AudioBox (Vyas et al., 2023) and SpeechCraft (Jin et al., 2024) proposed scaling up style-prompted TTS to a much larger pool of data; while Parler-TTS and SpeechCraft did so solely using automatic 430 tagging pipelines, AudioBox used a combination of automatically tagged data and internally anno-431 tated stylistic datasets to train the model. LibriTTS-P (Kawamura et al., 2024) explores abstract intrinsic tag annotations by annotating speakers in LibriTTS-R (Koizumi et al., 2023). Other relevant, contemporaneous style-prompted TTS work includes Chen et al. (2024); Zhu et al. (2024); Yamamoto et al. (2024).

Style Control for other Speech Tasks Recent work has explored natural language style prompts for tasks other than TTS. DreamVoice (Hai et al., 2024), like LibriTTS-P, annotates LibriTTS-R with abstract intrinsic tags, but for the task of voice conversion. The contemporaneous VCTK-RVA (Sheng et al., 2024) annotates the VCTK dataset with intrinsic tags for training a speech editing system that conditions on a style prompt instruction.

### 7 CONCLUSION

We propose a crisper definition of speech style tags, categorizing into abstract intrinsic, abstract situational, demographic and automatic tags. We use this to substantially expand the space of style prompts by supporting 63 total tags. Emphasizing the importance of abstract tags, we collect intrin-sic tag human annotations for a subset of speakers in the VoxCeleb (Nagrani et al., 2020) dataset to create StyledVoxCeleb, and reuse Expresso (Nguyen et al., 2023) and EARS (Richter et al., 2024) for situational tags. We train style-prompted TTS models based on Parler-TTS (Lacombe et al., 2024b; Lyth & King, 2024) that show improved performance on speech-style consistency metrics compared to competitive baselines, while they underperform baselines on speech quality and content correctness metrics.

### 8 LIMITATIONS

Expensive human annotation Our dataset collection strategy relies on expensive, slow human annotation for abstract intrinsic and situational tags. While it is significantly cheaper to annotate intrinsic tags on a speaker level rather than situational tags on an utterance level, it is unclear how to substantially and cheaply scale either type of annotation. Future work could potentially look into using synthetic data augmentation (Défossez et al., 2024) for automatically expanding existing annotated datasets.

**Noisy data** While VoxCeleb is beneficial as it has a high diversity of speakers and is sourced from a more realistic, in-the-wild speech domain, it negatively affects model performance when evaluated for speech quality and content correctness due to inherent background noise in a majority of its utterances. While this affects our models that are trained on a subset of VoxCeleb, we expect that scaling to more speakers will mitigate this issue to some extent.

**Language coverage** We limit our current experiments to English data; there is a lot of potential to expand style-prompted TTS to more languages, both in terms of the language of the utterance and the language of the style prompt. Some work (Jin et al., 2024; Yamamoto et al., 2024) explores other languages like Chinese and Japanese in addition to English for style-prompted TTS.

## References

- Zhiyong Chen, Xinnuo Li, Zhiqi Ai, and Shugong Xu. Stylefusion tts: Multimodal style-control and enhanced feature fusion for zero-shot text-to-speech synthesis, 2024. URL https://arxiv. org/abs/2409.15741.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models, 2022.

486 Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, 487 Edouard Grave, and Neil Zeghidour. Moshi: a speech-text foundation model for real-time dia-488 logue. Technical report, Kyutai, September 2024. URL http://kyutai.org/Moshi.pdf. 489 Sanchit Gandhi, Patrick von Platen, and Alexander M. Rush. Distil-whisper: Robust knowledge 490 distillation via large-scale pseudo labelling, 2023. URL https://arxiv.org/abs/2311. 491 00430. 492 493 Wenhao Guan, Yishuang Li, Tao Li, Hukai Huang, Feng Wang, Jiayan Lin, Lingyan Huang, Lin 494 Li, and Qingyang Hong. Mm-tts: Multi-modal prompt based style transfer for expressive text-to-495 speech synthesis, 2024. URL https://arxiv.org/abs/2312.10687. 496 497 Zhifang Guo, Yichong Leng, Yihan Wu, Sheng Zhao, and Xu Tan. Prompttts: Controllable text-tospeech with text descriptions, 2022. URL https://arxiv.org/abs/2211.12171. 498 499 Jiarui Hai, Karan Thakkar, Helin Wang, Zengyi Qin, and Mounya Elhilali. Dreamvoice: Text-guided 500 voice conversion, 2024. URL https://arxiv.org/abs/2406.16314. 501 502 Shengpeng Ji, Jialong Zuo, Minghui Fang, Ziyue Jiang, Feiyang Chen, Xinyu Duan, Baoxing Huai, and Zhou Zhao. Textrolspeech: A text style control speech corpus with codec language text-to-504 speech models. In ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and 505 Signal Processing (ICASSP). IEEE, April 2024. doi: 10.1109/icassp48485.2024.10445879. URL http://dx.doi.org/10.1109/ICASSP48485.2024.10445879. 506 507 Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chap-508 lot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, 509 Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, 510 Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7b, 2023. URL https: 511 //arxiv.org/abs/2310.06825. 512 513 Zeyu Jin, Jia Jia, Qixin Wang, Kehan Li, Shuoyi Zhou, Songtao Zhou, Xiaoyu Qin, and Zhiyong Wu. 514 Speechcraft: A fine-grained expressive speech dataset with natural language description. In ACM 515 Multimedia 2024, 2024. URL https://openreview.net/forum?id=rjAY1DGUWC. 516 Masaya Kawamura, Ryuichi Yamamoto, Yuma Shirahata, Takuya Hasumi, and Kentaro Tachibana. 517 Libritts-p: A corpus with speaking style and speaker identity prompts for text-to-speech and style 518 captioning, 2024. URL https://arxiv.org/abs/2406.07969. 519 Yuma Koizumi, Heiga Zen, Shigeki Karita, Yifan Ding, Kohei Yatabe, Nobuyuki Morioka, Michiel 521 Bacchiani, Yu Zhang, Wei Han, and Ankur Bapna. Libritts-r: A restored multi-speaker text-tospeech corpus, 2023. URL https://arxiv.org/abs/2305.18802. 522 523 Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. High-524 fidelity audio compression with improved rvggan, 2023. URL https://arxiv.org/abs/ 525 2306.06546. 526 527 Yoach Lacombe, Vaibhav Srivastav, and Sanchit Gandhi. Data-speech. https://github.com/ 528 ylacombe/dataspeech, 2024a. 529 Yoach Lacombe, Vaibhav Srivastav, and Sanchit Gandhi. Parler-tts. https://github.com/ 530 huggingface/parler-tts, 2024b. 531 532 Marvin Lavechin, Marianne Métais, Hadrien Titeux, Alodie Boissonnet, Jade Copet, Morgane Rivière, Elika Bergelson, Alejandrina Cristia, Emmanuel Dupoux, and Hervé Bredin. Brouhaha: 534 multi-task training for voice activity detection, speech-to-noise ratio, and C50 room acoustics 535 estimation. ASRU, 2023. 536 Yichong Leng, Zhifang Guo, Kai Shen, Xu Tan, Zeqian Ju, Yanqing Liu, Yufei Liu, Dongchao Yang, Leying Zhang, Kaitao Song, Lei He, Xiang-Yang Li, Sheng Zhao, Tao Qin, and Jiang 538 Bian. Prompttts 2: Describing and generating voices with text prompt, 2023. URL https: //arxiv.org/abs/2309.02285.

548

554

556

- 540 Guanghou Liu, Yongmao Zhang, Yi Lei, Yunlin Chen, Rui Wang, Zhifei Li, and Lei Xie. Prompt-541 style: Controllable style transfer for text-to-speech with natural language descriptions, 2023. URL 542 https://arxiv.org/abs/2305.19522.
- Haohe Liu, Oiugiang Kong, Oiao Tian, Yan Zhao, DeLiang Wang, Chuanzeng Huang, and Yuxuan 544 Wang. Voicefixer: Toward general speech restoration with neural vocoder, 2021.
- 546 Dan Lyth and Simon King. Natural language guidance of high-fidelity text-to-speech with synthetic 547 annotations, 2024. URL https://arxiv.org/abs/2402.01912.
- Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, Shiliang Zhang, and Xie Chen. 549 emotion2vec: Self-supervised pre-training for speech emotion representation, 2023. URL 550 https://arxiv.org/abs/2312.15185. 551
- 552 Arsha Nagrani, Joon Son Chung, Weidi Xie, and Andrew Zisserman. Voxceleb: Large-scale speaker 553 verification in the wild. Computer Speech & Language, 60:101027, 2020.
- Tu Anh Nguyen, Wei-Ning Hsu, Antony D'Avirro, Bowen Shi, Itai Gat, Maryam Fazel-Zarani, Tal 555 Remez, Jade Copet, Gabriel Synnaeve, Michael Hassid, Felix Kreuk, Yossi Adi, and Emmanuel Dupoux. Expresso: A benchmark and analysis of discrete expressive speech resynthesis, 2023. URL https://arxiv.org/abs/2308.05725.
- OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Floren-559 cia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Moham-561 mad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, 564 Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, 565 Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey 566 Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, 567 Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila 568 Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gib-569 son, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan 570 Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hal-571 lacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan 572 Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, 573 Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun 574 Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Ka-575 mali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook 576 Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel 577 Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen 578 Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel 579 Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv 580 Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, 581 Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, 582 Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel 583 Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Ra-584 jeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, 585 Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, 588 Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, 589 Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, 592 Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang,

626

637

645

594 Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Pre-595 ston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vi-596 jayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan 597 Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, 598 Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao 600 Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL 601 https://arxiv.org/abs/2303.08774. 602

- Puyuan Peng, Po-Yao Huang, Shang-Wen Li, Abdelrahman Mohamed, and David Harwath. Voicecraft: Zero-shot speech editing and text-to-speech in the wild, 2024. URL https://arxiv. org/abs/2403.16973.
- Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. Mls:
   A large-scale multilingual dataset for speech research. In *Interspeech 2020*. ISCA, October 2020. doi: 10.21437/interspeech.2020-2826. URL http://dx.doi.org/10.21437/
   Interspeech.2020-2826.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022. URL https://arxiv. org/abs/2212.04356.
- Flávio Ribeiro, Dinei Florêncio, Cha Zhang, and Michael Seltzer. Crowdmos: An approach for crowdsourcing mean opinion score studies. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2416–2419, 2011. doi: 10.1109/ICASSP.2011.
  5946971.
- Julius Richter, Yi-Chiao Wu, Steven Krenn, Simon Welker, Bunlong Lay, Shinji Watanabe, Alexander Richard, and Timo Gerkmann. Ears: An anechoic fullband speech dataset benchmarked for speech enhancement and dereverberation, 2024. URL https://arxiv.org/abs/2406.06185.
- <sup>624</sup> Zhengyan Sheng, Yang Ai, Li-Juan Liu, Jia Pan, and Zhen-Hua Ling. Voice attribute editing with text prompt, 2024. URL https://arxiv.org/abs/2404.08857.
- 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. Audiobox: Unified
   audio generation with natural language prompts, 2023. URL https://arxiv.org/abs/2312.15821.
- Chengyi Wang, Sanyuan Chen, Yu Wu, Ziqiang Zhang, Long Zhou, Shujie Liu, Zhuo Chen, Yanqing
   Liu, Huaming Wang, Jinyu Li, Lei He, Sheng Zhao, and Furu Wei. Neural codec language models
   are zero-shot text to speech synthesizers, 2023. URL https://arxiv.org/abs/2301.
   02111.
- Aya Watanabe, Shinnosuke Takamichi, Yuki Saito, Wataru Nakata, Detai Xin, and Hiroshi Saruwatari. Coco-nut: Corpus of japanese utterance and voice characteristics description for prompt-based control. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–8. IEEE, 2023.
- Ryuichi Yamamoto, Yuma Shirahata, Masaya Kawamura, and Kentaro Tachibana. Description based controllable text-to-speech with cross-lingual voice control, 2024. URL https://
   arxiv.org/abs/2409.17452.
- Dongchao Yang, Songxiang Liu, Rongjie Huang, Chao Weng, and Helen Meng. Instructtts: Modelling expressive tts in discrete latent space with natural language style prompt, 2023. URL https://arxiv.org/abs/2301.13662.

 Kinfa Zhu, Wenjie Tian, Xinsheng Wang, Lei He, Yujia Xiao, Xi Wang, Xu Tan, sheng zhao, and Lei Xie. Unistyle: Unified style modeling for speaking style captioning and stylistic speech synthesis. In ACM Multimedia 2024, 2024. URL https://openreview.net/forum? id=7BZ4biy975.

A LIST OF SPEECH STYLE TAGS

655	TTL: 's due l'est a Constant a sous d'Alexa
656	This is the list of tags we consider:
657	• Intrinsic:
658	– Abstract:
659	* <b>Pitch</b> : Shrill Nasal Deen
660	* <b>Texture</b> Silky Husky Raspy Guttural Vocal-fry
661	* Clarity: Crisp Slurred Lisp Stammering
662	* <b>Volume:</b> Booming, Authoritative, Loud, Hushed, Soft.
663	* <b>Rhythm:</b> Pitchy, Flowing, Monotonous, Staccato, Punctuated, Hesitant,
664	Singsong, Enunciated.
665	– Demographic:
666	* Gender: Male, Female.
667	* Accent: American, British, Scottish, Canadian, Australian, Irish, Indian, Ja-
668	maican.
669	– Automatic:
670	* Pitch Levels: High-pitched, Medium-pitched, Low-pitched.
671	Situational:
672	– Abstract:
673	* <b>Emotion:</b> Enthusiastic, Happy, Angry, Saddened, Awed, Calm, Anxious, Dis-
674	gusted, Scared, Confused, Bored, Sleepy, Pained, Guilt, Sarcastic, Sympathetic,
675	Admiring, Desirous.
676	* <b>Expressiveness:</b> Animated, Laughing, Passive, Whispered.
677	– Automatic:
678	* Speaking Rate Levels: Fast, Measured, Slow.
679	
680	We note that our datasets contain more accent tags than those shown here, but these 8 accents are
681	most represented in our datasets, and hence we evaluate on only these accents (accent distribution of
682	our datasets can be viewed at Figure 2 some tag categories like volume, speaking rate and rhythm

<sup>681</sup> indicates in our datasets, and hence we evaluate on only these accents (accent distribution of our datasets can be viewed at Figure 2 Some tag categories like volume, speaking rate and rhythm can span both intrinsic and situational; however, we collect data for volume with intrinsic human annotations, and automatically obtain speaking rate tags on an utterance-by-utterance basis i.e. in a situational manner. Therefore, we place them in their respective intrinsic or situational categories. We collect all rhythm tags with intrinsic annotations, and place them in the intrinsic category; however, 2 rhythm tags (*singsong, enunciated* are also present in our situational datasets which we also use. The manually written definitions for each style tag can be found in Table 5.

688 689 690

691

692

693 694

695

652 653

654

#### **B** DATASET STATISTICS

We report dataset statistics for accent distribution in Figure 2 and abstract tags distribution in Figure 3.

C DATASET PREPROCESSING

For all datasets, we convert audio from their original format to the .wav format, apply loudness normalization using SoX and PyDub <sup>12</sup> such that the peak volume of each audio is -0.1 dB, and discard all audios shorter than 2 s or longer than 30 s. If an utterance does not come with ground truth transcripts, we synthesize transcripts using the Whisper (Radford et al., 2022) large-v3 ASR model. We describe dataset-specific preprocessing below:

<sup>&</sup>lt;sup>12</sup>https://sourceforge.net/projects/sox/, https://github.com/jiaaro/pydub

Attribute	Description
High-pitched	A voice with a distinctly high frequency.
Shrill	A high-pitched, piercing, and sharp voice.
Nasal Medium-pitched	A whiny voice that sounds like someone is speaking through their nose. A voice with a medium frequency that is neither very high or low-nitched
Low-pitched	A voice with a distinctly low frequency.
Deep	A low-pitched, resonant, rich voice.
Silky	A smooth, pleasant and soothingly soft voice.
Raspy	A singing rough, now voice that conveys a gritty texture. A rough, grating, somewhat harsh voice.
Guttural	A deep, throaty, gravelly voice.
Vocal-fry	A creaky, breathy voice that occurs when vocal cords flutter and produce a sizzling, popping sour
American	A voice with an American accent.
British	A voice with a British accent.
Scottish	A voice with a Scottish accent.
Canadian	A voice with a Canadian accent.
Irish	A voice with an Irish accent.
Indian	A voice with an Indian accent.
Jamaican	A voice with an Jamaican accent.
Male	A male voice, often having a lower pitch.
Booming	A loud, resonant, commanding, powerful voice.
Authoritative	A confident, clear voice with a tone that conveys expertise and assurance.
Loud	A voice with a high volume.
Hushed	A soft, quiet, low-volume voice typically used to convey intimacy or secrecy.
Whispered	A breathy, low-volume, cann and sooning voice typically used to convey sublicity.
Crisp	A clear, distinct, articulate voice.
Slurred	An unclear, difficult-to-understand voice that blends together sounds and words.
Lisp	with a 'th' sound.
Stammering	A voice with pauses, repetitions and prolongations of words that disrupt the speech flow.
Singsong	A melodious voice that rises and falls in a musical manner.
Pitchy	A jarring, somewhat unstable voice that often strays from the correct pitch.
Monotonous	A dull, flat voice whose pitch, tone and speed remains constant throughout.
Staccato	A disjointed, unclear voice with breaks in-between syllables or words.
Punctuated	An engaging voice with clear, deliberate pauses that emphasize key words.
Fast speed	A voice that clearly and precisely articulates words, with each synable distinctly pronounced. A rapidly speaking, quick voice with few pauses.
Measured speed	A controlled, deliberate voice that has an even tone and a moderate speed.
Slow speed	A voice with a slower speaking rate.
Hesitant Enthusiastic	An uncertain, tentative voice, often marking a lack of confidence, refluctance of confusion. A lively energetic positive voice that conveys excitement and interest in the topic being discusse
Нарру	A warm, positive and joyful voice.
Angry	A raised voice that conveys anger, frustration or displeasure, characterized by raised volume and
Coddonad	emphatic speech patterns.
Awed	A voice that conveys the speaker's admiration, wonder or reverance of something the speaker appr
	ciates.
Calm	A calm, gentle and serene voice that conveys the speaker's relaxed and peaceful emotion.
Anxious Disgusted	A voice that conveys nervousness and anxiety, often marked by rapid or jittery speech patterns. A intenated voice that conveys repulsion and disgust by appropriately altering its nitch and rbyth
Scared	A shaky, rapid voice that reflects the speaker's anxiety or fear.
Confused	A voice characterized by indecision and a lack of clarity, often marked by hesitance.
Bored	A voice, often monotonous, that indicates lack of enthusiasm and disinterest.
Pained	A voice characterized by a strained, trembling tone that indicates sorrow or anguish.
Guilt	A voice that carries a wavering, hesitant tone that hints at discomfort or regret.
Sarcastic	A speaking style that is characterized by a distinct tone of irony that suggests that the speaker
Sympathetic	Intention is to mock or convey contempt. A gentle, compassionate voice that reassures and seeks to empathize with the listener
Admiring	An appreciative, positive and complimentary manner of speaking.
Desirous	An emotional voice that conveys deep longing or desire.
Animated	A energetic, heightened voice characterized by varied intonations or emotional intensity.
Passive	A voice with intermittent sounds of laughter conveying amusement and joy. A tentative, subdued and uninterested voice.



#### C.1 STYLEDVOXCELEB

801

809

We combine the VoxCeleb1 and VoxCeleb2 datasets. We apply a noise removal model, Voicefixer (Liu et al., 2021) to all audios, since we observed that a significant proportion of VoxCeleb data is noisy (the median SNR for VoxCeleb data is 31.76 dB computed by Brouhaha (Lavechin et al., 2023); compare to 59.49, 50.42 and 61.70 for Expresso, EARS and LibriTTS-R respectively). We then run a language identification model Lingua <sup>13</sup> over the transcripts and only keep those examples whose transcripts are identified as English text and discard celebrities with fewer than 10 English audio clips.

<sup>&</sup>lt;sup>13</sup>https://github.com/pemistahl/lingua-py

#### C.2 EXPRESSO

The Expresso dataset consists of 4 voice actors speaking various utterances in different speaking styles. We discard the *default*, *narration* and *non-verbal* speaking styles, since they do not exhibit the situational tags we are interested in. Since some of the data is in the form of long freeform dual-channel conversations between two voice actors, we use the Voice Activity Detection data provided by the dataset to splice the long conversation into two channels and VAD-segmented chunks, so that we can use each chunk as an utterance. We then remap each speaking style to our tag vocabulary as described in Section C.4.

C.3 EARS

The EARS dataset consists of 107 speaking enacting various speaking styles. We discard the long freeform examples as they are not labelled with speaking styles. We also discard interjection, nonverbal and vegetative speaking styles since they do not contain natural speech. We remap the speaking styles in the rest of the data to our tag vocabulary as described in Section C.4.

#### TERM REMAPPING C.4

We remap terms in the Expresso, EARS, and LibriTTS-P datasets to terms in our vocabulary using the mapping in Table 6. 

835				
836	Original Term	Mapped Term(s)	Original Term	Mapped Term(s)
837	feminine	female	awe	awed
838	halting	stammering	bored	bored, passive
839	tensed	anxious	desire	desirous, animated
840	relaxed	calm	projected	loud
841	powerful	authoritative	fearful	scared
842	muffled	slurred	amusement	happy
843	masculine	male	distress	anxious, scared
844	fluent	flowing	disappointment	saddened, passive
9/5	weak	hushed	realization	awed
040	sharp	crisp	amazement	awed
040	reassuring	sympathetic	disgust	disgusted
847	lively	enthusiastic	fear	scared
848	cool	calm	anger	angry
849	happy	happy, animated	adoration	admiring
850	laughing	laughing, animated	confusion	confused
851	sad	saddened	interest	enthusiastic
852	whisper	whispered	serenity	calm
853	singing	singsong	contentment	calm, passive
854	angry	angry, animated	sadness	saddened
855	desire	desirous	extasy	happy
856	interest	enthusiastic	pain	pained
957	serenity	calm	cuteness	happy
057	contentment	calm, passive	relief	calm, passive
808	sadness	saddened	pride	admiring
859	loud	loud	embarrassment	anxious
860	whisper	whispered		
861				

Table 6: Terms in existing datasets remapped to terms in our vocabulary.

# B64 D LLM PROMPTING

866

```
D.1 IMPERFECTLY LABELLING CELEBRITIES WITH STYLE TAGS
We use the gpt-4-0125-preview version of GPT-4 via the OpenAI API with the default hyperparameters (temperature 1.0, top-p 1.0, maximum 2048 tokens). We prompt it with the name of
```

```
the celebrity and ask it to output a list of style tags associated with the celebrity's voice with the
870
       following prompt template:
871
872
       Given the name of a famous celebrity or actor, you must retrieve
873
       \rightarrow your knowledge about that celebrity's voice and map the voice
874
       \rightarrow to a subset of speech style attribute labels provided to you.
875
       \, \hookrightarrow \, Here is the list of speech style attribute types you should
876
       \rightarrow pay attention to, along with attribute labels for each type:
877
       <attributes>
878
       - **Pitch:** Shrill, Nasal, Deep.
879
       - **Texture:** Silky, Husky, Raspy, Guttural, Vocal-fry.
       - **Volume:** Booming, Authoritative, Loud, Hushed, Soft.
880
       - **Clarity:** Crisp, Slurred, Lisp, Stammering.
881
       - **Rhythm:** Singsong, Pitchy, Flowing, Monotonous, Staccato,
882
       → Punctuated, Enunciated, Hesitant.
883
       </attributes>
884
885
       Your task is to associate the celebrity with a subset of these
886
       \rightarrow attributes, taking into account how the celebrity always
887
       \, \hookrightarrow \, sounds like. Only use the attributes that are extremely
888
       \rightarrow salient to the celebrity's voice i.e. their unique speech
889
           styles. Don't create any new attributes because you will fail
       \hookrightarrow
890
           the task if you do so.
       \hookrightarrow
891
       The celebrity is {name}. First generate a paragraph of around 5
892
           sentences, within <description> tags, using your knowledge,
893
       \hookrightarrow
       \hookrightarrow that describes the salient attributes of {name}'s voice. Then,
894
       \, \hookrightarrow \, within <attribute> tags, generate a list of comma-separated
895
       \rightarrow speech style attributes, from the above attributes list, that
896
           saliently apply to {name}. Use the following format:
       \hookrightarrow
897
       <description>
       (Description goes here)
899
       </description>
900
       <attribute>
901
       (Comma-separated list of attributes)
902
       </attribute>
903
904
       D.2 EXTRACTING GENDER AND ACCENT
905
       We use the gpt-4-0125-preview version of GPT-4 via the OpenAI API with the default hyper-
906
       parameters (temperature 1.0, top-p 1.0, maximum 2048 tokens). We prompt it with the name of the
907
       celebrity and ask it to output the celebrity's gender and accent with the following prompt template:
908
909
       Tell me the accent and the gender of {name} formatted as
910
       Accent: <accent>
911
       Gender: <gender>
912
913
       D.3 GENERATING STYLE PROMPTS
914
```

We use the Mistral-7B-Instruct-v0.2 LLM (Jiang et al., 2023) to generate prompts via the Dataspeech
library with a per-device batch size of 32 and sample with a temperature of 0.6, a top-p of 1.0 with
a maximum 256 new tokens. We prompt the model with a comma-separated list of style tags and
instruct it to generate a style prompt with the following prompt:

918 An audio sample of a person's speech can be described in several 919 ways using descriptive keywords. These keywords may include 920 demographic data about the person (e.g. gender, name, accent)  $\hookrightarrow$ 921 and voice characteristics (e.g. related to pitch, gender,  $\hookrightarrow$ 922 texture and rhythm, volume, clarity, speaking rate, emotion,  $\hookrightarrow$ expressiveness). 923  $\hookrightarrow$ 924 You will be provided several keywords that describe the speech 925 sample. Your task is to create a simple text description using  $\hookrightarrow$ 926 the provided keywords that accurately describes the speech  $\hookrightarrow$ 927 sample. Ensure that the description remains grammatically  $\hookrightarrow$ 928 correct, easy to understand, and concise. You can rearrange  $\hookrightarrow$ 929 the keyword order as necessary, and substitute synonymous 930 terms where appropriate. After you are provided the keywords,  $\hookrightarrow$ 931  $\hookrightarrow$ generate only the description and do not output anything else. 932 933 An example is provided below. female, confused, hesitant, slightly noisy environment 934 935 Description: A woman's speech sounds confused and hesitant, 936 → recorded in a slightly noisy environment. 937 938 Now, generate a description for the following example: 939 {all\_tags\_str} 940 941 Description: 942 943 Е HUMAN ANNOTATION: DETAILS 944 945

946 E.1 ANNOTATION DETAILS

947 We recruit Amazon Mechanical Turk workers certified as Verified workers with a minimum approval 948 rate of 98% and at least 500 successful HITs. We perform a qualification task using 6 pairs of 949 manually selected clips from VoxCeleb or Expresso where one clip exhibits a style (one of *deep*, 950 whispered, scared, slurred, high-pitched, enunciated) and the other doesn't, and select those 38 951 annotators that succeed in finding the right clip for at least 5 of the 6 pairs. We use this pool 952 of annotators for our data collection. For evaluation metrics, we use all Verified workers with a minimum approval rate of 98% and at least 500 successful HITs rather than just our pool of 38 953 workers for faster evaluation turnaround. We pay annotators \$9/hr. 954

E.2 ANNOTATION USER INTERFACES

We display the annotation UIs for qualification task in Figure 4, crowdsourcing abstract intrinsic
style tag annotations in Figure 5, speech quality evaluation in Figure 6, and speech-style consistency
evaluation in Figure 7.

961 962

955

- 963
- 964
- 965
- 966
- 967 968
- 969
- 970
- 971

Instructions		
Welcome to our speech style attribute evaluation task! Here are instructions on how to use this interface:	r'e voice et de in each alle	
2. For any presented with two speech clips below. Listen to both clips, paying careful attention to the speake 2. Below the speech clips, you are asked to select which clip better matches the specified style attribute. The	style attribute is a speech characteristic lik	ke 'Deep', 'Whispered', 'Angry', etc. A description of w
attribute is available to better understand what the style means. Compare the two clips and select the one th clips, in which case you can select 'Neither'. If you think both clips equally fit the style attribute well and can	nat you think better fits the style attribute. not decide between them, you can select 'E	Sometimes, the style attribute may be completely ab 30th'.
3. Once you have made your choice, you can click the 'Save and Continue' button to save it and move to the	next annotation example. Wait for both clip	os to fully load.
4. Once you have completed all the audio clips, you will see a completion message with a survey code. Pleas 5. You can track how many examples you have apportated using the progress information right above the sport	e copy this code back to the Amazon Mech	anical Turk task to receive your payment.
5. Tou can u ack now many examples you have annotated using the progress mormation right above the spe FAO:	een cups.	
2: Should we pay attention to the voice style or the content of the speech? A: You should mainly focus on the voice style to make your decision.		
Q: What if there are multiple speakers or background noise in the clip?		
A: There should be only one primary speaker in the clip, although there may be background noise or a few so	conds where you hear other speakers. Plea	ase focus on the primary speaker's voice characteris
vote that you can cottapse these instructions by circking on the instructions text at the top.		
gress: Annotation 1 or 6.		
ана с , оЩ , фр. сали со настик и и илини	http://www.	In Il andie de erstree Incare
	- IIIIneeneer IIIne IIIIIIIIIIIIIIIIIIIII	
0:00	0:00	
	xI (()	≪ ♦ ₩
Which clip matches the style attribute 'Deep' better?	Style Attribute Info	
Clip 1 Clip 2 Neither Both	Style Attribute: Deep	-hunica
	Description: A tow-pitched, resonant, rid	un voice.
Save and	l Continue	
Figure 4: Annotation UI for	selecting qualified	annotators.
Instructions Welcome to our speech type annotation task Here are instructions on how to use this interface: 1. You so properties with a speech citle before, consisting of recordings of a single speaker. The name of the the the taskou below, based on what you heard, pleases type out at least 3 distinct speech single attributes, sep	speaker is provided. Please listen to the clip, paying co	refut attention to the speaker's voice characteristics. In speaker's voice.
Instructions Wetcome to our speech style annotation task Here are instructions on how to use this interface: 1. You are presented with a speech tighe bandwate to the state of the the testbox below, based on what you heard, pleases type out at least 3 distinct upsech shiple attributes, says 2. Once you have completed all the audio clips, you will see a completion message with a survey code. Pleas	speaker is provided. Please listen to the clip, paying ca rated by commas, that you think uniquely describe the ns and move no the next adulo clip. se corpy this code back to the Amazon Mechanical Turk.	reful attention to the speaker's voice characteristics. In speaker's voice.
Instructions Welcome to our speech style annotation task! Here are instructions on how to use this interface: 1. You are presented with a speech clip below, consisting of recordings of a single speaker. The name of the the testbox below, based on what you hand, plases type out at least 1 distinct appendix hyle atthibutes, seps 2. Once you have entered your answer, our can lick the Save and Continue Duttons to survey oroundix 3. Once you have entered your answer, our can lick the Save and Continue Duttons to survey oroundix 4. You can track how many examples you have annotated using the progress information right above the sg FAQ;	speaker is provided. Please listen to the clip, paying ca rated by commas, that you think uniquely describe the mand move no the next audio clip. se copy this code back to the Amazon Mechanical Turk eech clips.	reful attention to the speaker's voice characteristics. In speaker's voice. task to receive your payment.
Instructions Welcome to our speech style annotation takid Here are instructions on how to use this interface: 1. You are presented with a speech clip below, consisting of recordings of a single speaker. The name of the the thethose below, based on what you heard, pleases type out at least 3 distinct agees high efficiency. 2. Once you have entered your answers, you can lick the "save and continue" buttoms to save your anotation 3. Once you have entered your answers, you can lick the "save and continue" buttoms to save your anotation 3. Once you have entered your answers, you can lick the "save and continue" buttoms to save your anotation 3. Once you have entered your answers, you can lick the "save and continue" buttoms to save your anotate 4. You can track how many examples you have annotated using the progress information right above the sp <b>Fix</b> ; <b>O</b> : <b>O</b> : What (Ifther are multiple speakers or background noise in the clip?)	speaker is provided. Please listen to the clip, paying ca nated by commas, that you think uniquely describe the as done was the next addo clip. se copy this code back to the Amazon Mechanical Turk eech clip.	reful attention to the speaker's voice characteristics. In speaker's voice. task to receive your payment.
Instructions  Instructions  Welcome to our speech style annotation task! Here are instructions on how to use this interface:  I. You are presented with a speech clipb below, considing of recording of a single speaker. The name of the the thetable below, based on what you head, please type our at least 3 distinct speech style attributes, sep 2. Once you have completed all the audo clips, you will see a completion message with a survey code. The 4. You can track how many examples you have annotated using the progress information right above the ge <b>PQ</b> Q. Mult of there are multiple speakers or background noise in the clip?  A. These should be only one primary updater in the clips. Although the prior the background noise or a lew?  A. Wut of the speaker's voice change during the tcip?	speaker is provided. Please listen to the clip, paying ca rated by commas, that you think uniquely describe the ns and move on to the next audio clip. ee copy this code back to the Amazon Mechanical Turk eech clips.	reful attention to the speaker's voice characteristics. In speaker's voice. task to receive your payment. In the primary speaker's voice characteristics.
Instructions  Metcome to our speech style annotation task Here are instructions on how to use this interface:  1. You are presented with a speech dig below, consisting of recordings of a single speaker. The name of the the testbale below, based on what you hand, places types out a least 3 slinking tapes the single speaker. The name of the the testbale below, based on what you hand, places types out a least 3 slinking tapes the single speaker. The name of the the testbale below, based on what you hand, places types out a least 3 slinking tapes the single structures, up a 1. Once you have completed all the audio clips, you will see a completion message with a survey code. Plate 4. You can track how many seamilies you have annotated using the progress information right above the sg EQ: 2. What if there are multiple possibles or background noise in the clip? 3. Now should focus on the basic vice characteristics that are present in most of the recordings in the clips. 3. Note that you can collapse these instructions by clicking on the 'instruction' test at the top.	speaker is provided. Please listen to the clip, paying co rated by commas, that you think uniquely describe the ns and move on to the next audio clip. ee copy this code back to the Amazon Mechanical Turk eech clips. eeconds where you hear other speakers. Please focus of the basic characteristics of the speaker's voice should r	Influctations.
Instructions Metadome to our speech typic annotation task Here are instructions on how to use this interface: 1. You say represend with a speech clip before, considing of recordings of a single speaker. The name of the the techoic below, based on what you heard, please type out at least 3 distinct speech style attributes, saps 2. Once you have entered your answers, you can click the Save and Continue button to save your annotation 3. Once you have entered your answers, you can click the Save and Continue button to save your annotatio 3. Once you have entered your answers, you can click the Save and Continue button to save your annotatio 3. Once you have entered your answers, you can click the Save and Continue button to save your annotatio 3. Once you have entered your answers, you can click the Save and Continue button to save your annotatio 3. Once you have entered your answers, you can click the Save and Continue button to save your annotatio 3. Once you have entered your answers, you can click the Save and Continue button to save your annotatio 3. Once you have entered your answers, you can click the Save and Continue button to save your annotatio 3. Once you have entered your answers, you can click the Save and Continue button to save your annotatio 3. Once you have completed all the and click, you Wile set its asser you can be background noise or a few: 3. What (The speaker's work click characteristics that are present in most of the recordings in the clip): 3. What (The speaker's work click characteristics that are present in most of the recordings in the clip): 3. What (The speaker's work click characteristics that are present in most of the recordings in the clip): 3. What (The speaker's work click characteristics that are present in most of the recordings in the clip): 3. What (The speaker's work click characteristics that are present in most of the recordings in the clip): 3. What (The speaker's work click characteristics that are present in most of the recordings in the clip): 3. What (The speaker's	speaker is provided. Please listen to the clip, paying ca rated by commas, that you think uniquely describe the ns and move no to the next audio clip. se copy this code back to the Amazon Mechanical Turk eech clips. ecconds where you hear other speakers. Please focus o the basic characteristics of the speaker's voice should r	Infirite the speaker's voice characteristics. In speaker's voice.
Figure 4: Annotation UI for Instructions Metacome to our speech type annotation task Here are instructions on how to use this interface: 1. vous or presend with a speech clip biolog, considing of recordings of a single speaker. The name of the the testbox blow, based on what you bendy, pleases type out a least 3 distinct speech xige attributions, spee 2. Once you have completed all the assumption of the Save and Continuer buttom to save your annotation 3. Once you have completed all the assumption busines with a survey cole. The 4. vou can track how many examples you have annotated using the progress information right above the spee- <i>Distinct</i> 9. What if the speakers or background noise in the clip? At this should be one in one primary speaker in the clip, although there may be background noise or a few: 3. Vous due does on one basiar vision characteristics that are present in most of the recordings in the clip: At the should be out, one charges during the clip? At the should becan on the basiar vision characteristics that are present in most of the recordings in the clip: Note that you can collapse these instructions by clicking on the 'Instructions' test at the top. Progress: Annotation 1 dato: Progress: Annotation 1 dato:	speaker is provided. Please listen to the clip, paying ca rated by commas, that you think uniquely describe the ns and move on to the next audio clip. see copy this code back to the Amazon Mechanical Turk eech clips. econds where you hear other speakers. Please focus or the basic characteristics of the speaker's voice should r	reful attention to the speaker's voice characteristics. In speaker's voice. task to receive your payment. In the primary speaker's voice characteristics. not change much during the clip.
Figure 4: Annotation UI for Instructions Meteore to our speech style annotation task Here are instructions on how to use this interface: 1. vous on presented with a speech clip below, consisting of recordings of a single speaker. The name of the the the tactors below, based on what you heard, pleases type out at least 3 distinct speech single attributes, spei 2. Once you have completed all the saido clips, you will see a completion message with a survey code. They 3. Use can track how many examples you have annotated using the progress information right above the sp <i>Pige</i> 3. What if there are multiple speakers or background noise in the clip? 3. What if there are multiple speakers or background noise in the clip? 3. What if there are multiple speakers or background noise in the clip? 3. What if there are multiple speakers or background noise in the clip? 3. What if there are multiple speakers or background noise in the clip? 3. What if there are multiple speakers or background noise in the clip? 3. What if there are multiple speakers or background noise or the tecordings in the clip? 3. What if there are multiple speakers or background noise in the clip? 3. What if there are multiple speakers or background noise or the tecordings in the clip? 3. What if there are multiple speakers or background noise or the tecordings in the clip? 3. What if there are multiple speakers or background noise or the tecordings in the clip? 3. What if there are multiple speakers or background noise or the tecordings in the clip? 3. You should focus on the basic voice characteristics that are present in most of the recordings in the clip? 3. You should focus on the basic voice characteristics that are present in most of the recordings in the clip? 3. You should focus on the basic voice characteristics that are present in most of the recording in the clip? 3. You should focus on the basic voice characteristics that are present in most of the recording in the clip? 3. You should focus on the basic voice charact	speaker is provided. Please listen to the clip, paying ca rated by commas, that you think uniquely describe the ns and move on to the next audio clip. ex copy this code back to the Amazon Mechanical Turk eech clips. econds where you hear other speakers. Please focus o the basic characteristics of the speaker's voice should r	Infortations.
Figure 4: Annotation UI for  Instructions  Meteore to our speech style annotation task Here are instructions on how to use this interface:  Now any research with a speech citip below, consisting of recordings of a single speaker. The name of the the testbox below, based on what you heard, pleases type out a test 3 distinct speech style attributes, spe 2. Once you have completed all the audio citip, you will see a completion message with a survey code. Pile 3. Once you have completed all the audio citip, you will see a completion message with a survey code. Pile 3. Once you have completed all the audio citip, you will see a completion message with a survey code. Pile 3. Once you have completed all the audio citip, you will see a completion message with a survey code. Pile 3. Once you have completed all the audio citip, you will see a completion message with a survey code. Pile 3. Once you have completed all the audio citip, you will see a completion message with a survey code. Pile 3. Once you have completed all speech or the chaps. 3. Once you have completed all the audio citip, you will see a completion message with a survey code. Pile 3. These should be only one primary yapeaier in the citip, it. 3. When the these appeards you charge during the citip 3. You should focus on the basic voice characteristics that are present in most of the recordings in the citip. 3. Note should focus on the basic voice characteristics that are present in most of the recordings in the citip. 3. Note should focus on the basic voice characteristics that are present in most of the recordings in the citip. 3. Deserver. Amountation 1 d 107. 3. Deserver. 3. Deserve	speaker is provided. Please listen to the clip, paying ca rated by commas, that you think uniquely describe the ns and move on to the next audio clip. exe cory this code back to the Amazon Mechanical Turk eech clips. econds where you hear other speakers. Please focus o the basic characteristics of the speaker's voice should r	Immotations.         reful attention to the speaker's voice characteristics. In speaker's voice.         speaker's voice.         In the primary speaker's voice characteristics.         not change much during the clip.         ()), (), (), (), (), (), (), (), (), (),
Figure 4: Annotation UI for Instructions Metacome to our speech style annotation task Here are instructions on how to use this interface: 1. You are presented with a speech clip below, consisting of recording of a single speaker. The name of the the the tabob bolicy, based on what you hard, places type out a least 3 distinct speech shyle attributes, spei 2. Once you have completed all the audio clips, you will see a completion message with a survey code. Plete 3. Once you have completed all the audio clips, you will see a completion message with a survey code. Plete 4. You can track how many examples you have annotated using the progress information right above the sp <b>Fig</b> 9. Will if the area multiple speaker or back ground noise in the clip? A You should focus on the basic voice characteristics that are present in most of the recordings in the clip. Note that you can collapse these instructions by clicking on the 'instruction' text at the top. The grees. Amondation 1 dist? <b>Desters: Amondation</b> 1 dist? <b>Dester: Amy Schumer</b>	speaker is provided. Please listen to the clip, paying ca rated by commas, that you think uniquely describe the ns and move on to the next audio clip. ec cords to code back to the Amazon Mechanical Turk eech clips. econds where you hear other speakers. Please focus o the basic characteristics of the speaker's voice should r	Inflict LIOTS.
Figure 4: Annotation Ul for Instructions Metacome to our speech style annotation task free are instructions on how to use this interface. 1. You are presented with a speech clip below, consisting of recordings of a single speaker. The name of the the that tools our speech style annotation task free are instructions on how to use this interface. 1. You are presented with a speech clip below, consisting of recordings of a single speaker. The name of the the that tools observe completed all the audio clips, you will see a completion message with a survey coule. Price 3. Once you have completed all the audio clips, you will see a completion message with a survey coule. Price 4. You can took how may examples you have annotated using the progress information right above the sp <i>Figu</i> 9. Will the speaker sy wile change during the clip? 1. You should focus on the basic voice characteristics that are present in most of the recordings in the clip: Note that you can collapse these instructions by clicking on the 'instructions' text at the too. The gress: Annotation 1 at 197. Expers: Annotation 1 at 197. Expers: Annotation 1 at 197. 1. Completed all the speaker spine change during the clip? 1. One speaker spine change during the clips of the spine of the	speaker is provided. Please listen to the clip, paying ca nated by commas, that you think unquery describe the nated by commas, that you think unquery describe the second the next addio clip. se copy this code back to the Amazon Mechanical Turk eeconds where you hear other speakers. Please focus o the basic characteristics of the speaker's voice should r the basic characteristics of the speaker's voice should r the basic characteristics of the speaker's voice should r	Inflict atoms.
Figure 4: Annotation UI for Instructions Methods on speech style annotation task Here are instructions on how to use this interface. 1. Now you have completed with a speech clip below, consisting of recordings of a single speaker. The name of the the totato below, based on what you have, places type out a test 3 distinct speech style attributes, spie 3. Once you have entered you answer, you can lick the Save and continue buttons uses you wonted to the set of the speaker with a survey code. Place 3. Once you have entered you answer, you can lick the Save and continue button to use you annotated using the progress information right above the speech style attributes pairs on taskes and continue button to save your monted of the speaker in the clip. Now and place the speech style attributes on you have annotated using the progress information right above the speech style attributes to the only one primary speaker in the clip. Now the speech style attributes the speech style attributes the speech style attributes the speech style attributes that you can collapse these instructions by clicking on the Instruction's test at the top. What if the speaker a weight speech style attributes that are present in most of the recordings in the clip. Now should focus on the basic voice characteristic that are present in most of the recordings in the clip. Note: Annotation 1 div. Tegers: Annotation 1 div. Description: The speaker attributes the speech style attributes. Description: The speaker style style style style style style style style attributes and style	speaker is provided. Please listen to the clip, paying co atated by commas, that you think uniquely discribe the asted by commas, that you think uniquely discribe the asted by commas, that you think uniquely discribe the second the next addio clip. econds where you hear other speakers. Please focus o the basic characteristics of the speaker's voice should r ===================================	Infirit CLEUTS.
Figure 4: Annotation UI for Instructions Medicane to car speech style annotation task! Here are instructions on how to use this interface: 1. You are presented with a speech cilp below, consisting of recordings of a single speaker. The name of the the thetabo below, based on what you hand, places type out a least 1 distinct appeal. The name of the the thetabo below, based on what you hand, places type out a least 1 distinct appeal. The name of the the thetabo below, based on what you hand, places type out a least 1 distinct appeal. The name of the 1. Once you have completed all the audic clips, you will exe a completion message with a survey code. Here 2. Once you have completed all the audic clips, you will exe a completion message with a survey code. Here 3. Once you have completed all the audic clips, you will exe a completion message with a survey code. Here 3. Once you have completed all the audic clips, you will exe a completion message with a survey code. Here 3. Once you have completed papeaker in the clips. How they the the recordings in the clips. 3. What if there are multiple speakers or background noise in the clips. 3. What if the speaker's voice characteristics that are present in most of the recordings in the clips. 3. What if the speaker's voice characteristics that are present in most of the recordings in the clips. 3. What if the speaker's voice characteristics that are present in most of the recordings in the clips. 3. What is the you can collapse these instructions by clicking on the 'Instructions' text at the top. 3. Exercise Annotation 1 al 197. 3. Exercise Annotation 1 a	speaker is provided. Please listen to the clip, paying contracted by commas, that you think uniquely describe the as and move on to the next audio clip. se copy this code back to the Amazon Mechanical Turk each clips. econds where you hear other speakers. Please focus of the basic characteristics of the speaker's voice should reflect the speaker's voice should r	Infirit CLEUTS.
Instructions         Welcome to car speech style amodation task! Here are instructions on how to use this interface:         1. vus are presented with a speech cipbe book, consisting of recordings of a single speaker. The name of the the thethot book, based is what you hard, places type out at least 3 distinct speech sityle attributes, sept 2. Once you have enrethed vul an speech style attributes, weigh 2. Once you have enrethed vul an speech style attributes, sept 3. Once you have enrethed vul an speech style attributes weigh 2. Once you have enrethed vul an speech style attributes.         9. Once you have completed at the audio clips, you will exe a completion message with a survey code. Plee 7. Note:         9. What (Hore are multiple speaker or background noise in the clip)?         9. What (Hore are multiple speaker or background noise in the clip)?         9. What (Hore are multiple speaker or background noise in the clip)?         9. What (Hore are multiple speaker to background noise in the clip)?         9. What (Hore are multiple speaker to background noise in the clip)?         9. What (Hore are multiple speaker to background noise or a few:         9. What (Hore are multiple speaker to background noise or a few:         9. What (Hore are multiple speaker) species that specers in most of the recordings in the clip?         1. What (Hore are multiple speaker) species that specers in most of the recordings in the clip?         1. What (Hore are multiple speaker) species that specers in most of the recordings in the clip?         1. Specer: Arry Schuters         1. Specer: Arry Schuters <td>speaker is provided. Please listen to the clip, paying carried by commas, that you think uniquely describe the nes and move on to the nest audio clip.         nes and move on to the nest audio clip.         accords where you hear other speakers. Please focus of the speaker's voice should r         minimum clip.         minim</td> <td>Infirit Clauders.</td>	speaker is provided. Please listen to the clip, paying carried by commas, that you think uniquely describe the nes and move on to the nest audio clip.         nes and move on to the nest audio clip.         accords where you hear other speakers. Please focus of the speaker's voice should r         minimum clip.         minim	Infirit Clauders.
Figure 4: Annotation UI for Instructions Wetcome to our speech style annotation task! Here are instructions on how to use this interface: 1. You speech style annotation task! Here are instructions on how to use this interface. 1. You speech style annotation task! Here are instructions on how to use this interface. 1. You speech style annotation task! Here are instructions on how to use this interface. 1. You speech style annotation task! Here are instructions on how to use this interface. 1. You speech style annotation task? You will see a compileto nessage with a survey cond. File 2. Once you have compileted all the audic clips, you will see a compileto nessage with a survey cond. File 3. Once you have compileted all the audic clips, you will see a compileto nessage with a survey cond. File 3. Once you have compilete posters or background noise in the clip? 3. What if there are multiple speaker or background noise in the clip? 3. What if there are multiple speaker or background noise in the clip? 3. You should locus on the balic voice characteristics that are present in most of the recordings in the clip? 3. You should locus on the balic voice characteristics that are present in most of the recordings in the clip? 3. You should locus on the balic voice characteristics that are present in most of the recordings in the clip? 3. You should locus on the balic voice characteristics that are present in most of the recordings in the clip? 3. You should locus on the balic voice characteristics that are present in most of the recordings in the clip? 3. You should locus on the balic voice characteristics that are present in most of the recordings in the clip? 3. You should locus on the balic voice characteristics that are present in most of the recordings in the clip? 3. You should locus on the balic voice characteristics that are present in most of the recordings in the clip? <td>speaker is provided. Please listen to the clip, paying carried by commas, that you think uniquely describe the ns and move on to the next audio clip. econds where you hear other speakers. Please focus of the speaker's voice should r in the basic characteristics of the speaker's voice should r the basic characteristics of the spea</td> <td>The to use other descriptive words.  A minimum and the invest.  A minimum</td>	speaker is provided. Please listen to the clip, paying carried by commas, that you think uniquely describe the ns and move on to the next audio clip. econds where you hear other speakers. Please focus of the speaker's voice should r in the basic characteristics of the speaker's voice should r the basic characteristics of the spea	The to use other descriptive words.  A minimum and the invest.  A minimum
Figure 4: Annotation Ul for Instructions We donne to our speech style annotation task Here are instructions on how to use this interface. 1. You are presented with a speech cipbe book, consisting of recordings of a single speaker. The name of the the thetaboo blow, based on what you hand, please type out a test 3 distinct speech right eatmothers, spea 2. Once you have completed all the audio clips, you will see a completion message with a survey our click the 3. You are track how many examples you will see a completion message with a survey our click the 3. You are track how many examples you will see a completion message with a survey our click the 3. You what of there are multiple speakers or background noise in the clips? 3. You should be only one primery speaker in the clips Mithough there mays be background noise or a fewr 3. You should be only one primery speaker in the clips Mithough there most of the recordings in the clips? 3. You should be only one primery speaker in the clips Mithough there mays be background noise or a fewr 3. You should be only one primery speaker in the clips Mithough there mays be background noise or a fewr 3. You should be only one primery speaker in the clips Mithough there most of the recordings in the clips? 3. You should be only one primery speaker in the clips Mithough there most of the recordings in the clips? 3. You should be only one primery speaker in the clips Mithough there most of the recordings in the clips? 3. You should be only one primery speaker in the clips Mithough there most of the recordings in the clips? 3. You should be only one primer speaker in the clips Mithough there most of the recordings in the clips? 3. You should be only one primer speaker so the speaker so there are the tops. 3. Decret Style Attributes with Definition 3. Decret Style Mithough there are the speaker so the speaker	speaker is provided. Please listen to the clip, paying carried by commas, that you think uniquely describe the ns and move on to the next audio clip. ac only this code back to the Amazon Mechanical Turk each clips. aconds where you hear other speakers. Please focus of the basic characteristics of the speaker's voice should r the basic characteristics of the speaker attributes; please feel between the basic characteristics of the speaker's voice should r the basic characteristics of the speaker's voice should r the basic characteristics of the speaker's voice should r the basic characteristics of the speaker's voice shoul	The to use other descriptive words.  A speaking through their nose.  Is geneking through their nose.  Is a protocol of the speaker and the sp
<section-header><section-header><section-header><section-header><section-header><text><text><list-item><list-item><list-item><text><text><text><text><text><text><text><text><text></text></text></text></text></text></text></text></text></text></list-item></list-item></list-item></text></text></section-header></section-header></section-header></section-header></section-header>	speaker is provided. Please listen to the clip, paying ca praked by commast, that you think uniquely describe the ns and move on to the next audio clip. ac only this code back to the Amazon Mechanical Turk een chips. aconds where you hear other speakers. Please focus of the basic characteristics of the speaker's voice should r th	The to use other descriptive words.  The to use other descriptive words.  The speaking through their nose.  The primary speaker's voice characteristics.  The p
<section-header><section-header><section-header><section-header><section-header><text><text><list-item><list-item><list-item><section-header><text><text><text><list-item><text><text><text><text><text><text></text></text></text></text></text></text></list-item></text></text></text></section-header></list-item></list-item></list-item></text></text></section-header></section-header></section-header></section-header></section-header>	Scieccurring quantification  speaker is provided. Please listen to the clip, paying ca rated by commas, that you think uniquely describe the as and move on to the next audio clip.  ac corp this code back to the Amazon Mechanical Turk ac corp this code back to the Amazon Mechanical Turk accords where you hear other speakers. Please focus of the basic characteristics of the speaker's voice should r  methods and the static structure of the speaker's voice should r  static characteristics of the speaker's voice should r  static cha	The to use other descriptive words.  Speaking through their nose.  Speaking their nose.  Speaking through their nose.  Speaki
Figure 4: Annotation Ul for Instructions We does not part of the annotation task Here are instructions on how to use this interface. 1. Nore you have completed with a speech clip below, consisting of recording of a single speaker. The name of the the taskeds below, based on what you have, alphate tasked with a speech clip below. 1. Once you have completed all the addo clips, you will see a completion message with a survey code. PRef B: 2000 and	speaker is provided. Please listen to the clip, paying carated by commas, that you think uniquely describe the snand move on to the next audio clip.         see corp this code back to the Amazon Mechanical Turk each clips.         accords where you hear other speakers. Please focus of the basic characteristics of the speaker's voice should r         the basic characteristics of the speaker's voice should r         b basic characteristics of the speaker's voice should r         b basic characteristics of the speaker's voice should r         b basic characteristics of the speaker's voice should r         b basic characteristics of the speaker's voice should r         b basic characteristics of the speaker's voice should r         b basic characteristics of the speaker's voice should r         b basic characteristics of the speaker's voice should r         b so only a small set of possible attributes; please feel         Definitions (scroll to see more)         b Shift. A hight price di prioring and sharp voice.         b Shift. A hight price di prioring and sharp voice.         b Shift. A hight prioring how voice that comps ja Shift or noome ja Shift. A hight price di prioring and sharp voice.         b Shift. A hight price di prioring and sharp voice.         b Shift. A hight price di prioring and sharp voice.         b Shift. A hight price di prioring and sharp voice.         b Shift. A hight price di prioring and sharp voice.         b Shift. A hight prioring how voice	The to use other descriptive words.  Sequenting through their nose.  Sequenting through their nose.  Sequenting the clip.  The to use other descriptive words.  Sequenting the clip.  The to use other descriptive words.  Sequenting through their nose.  Se
Instructions   Metacons to use speech diple annotation task Here are instructions on hour to use this instructions   1. No use presented with a speech diple annotation task Here are instructions on hour to use this instructions   1. Once you have completed all the audio dips, you will see a completion message with a survey code. RHe is a completion message with a survey code. RHe is a completion message with a survey code. RHe is a completion message with a survey code. RHe is a completion message with a survey code. RHe is a completion message with a survey code. RHe is a completion message with a survey code. RHe is a completion message with a survey code. RHe is a survey code is discussion message with a survey code. RHe is a survey code is discussion message with a survey code. RHe is a survey code is discussion message with a survey code. RHe is a survey code is discussion message with a survey code. RHe is a survey code is discussion message with a survey code. RHe is a survey code is discussion message with a survey code. RHe is a survey code is discussion message with a survey code. RHe is a survey code is discussion message with a survey code. RHe is a survey code is discussion message with a survey code. RHe is a survey code is discussion message with a survey code. RHe is a survey code is discussion of the isolation of the discussion of the discussion of the isolation	speaker is provided. Please listen to the clip, apping carated by commas, that you think uniquely describe the sn and move no to the next audio clip.         see corp this code back to the Amazon Mechanical Turk each clip.         see corp this code back to the Amazon Mechanical Turk each clip.         excords where you hear other speakers. Please focus of the basic characteristics of the speaker's voice should r         minimum file	Information of the speaker's voice characteristics. In speaker's voice characteristics. In speaker's voice characteristics. In the primary speaker's voice characteristics. I
<section-header><section-header><section-header><section-header><section-header><section-header><section-header><section-header><section-header><section-header><section-header><section-header><text><text><list-item><list-item><list-item><text><text><text><text><text></text></text></text></text></text></list-item></list-item></list-item></text></text></section-header></section-header></section-header></section-header></section-header></section-header></section-header></section-header></section-header></section-header></section-header></section-header>	Scieccuring quantified  speaker is provided. Please listen to the clip, paying ca rated by commas, that you think uniquely describe th  sa and move no to the next audio clip.  see cory this code back to the Amazon Mechanical Turk each clips.  conds where you hear other speakers. Please focus o  the basic characteristics of the speaker's voice should r  method of the speaker's voice should r  method of the speaker's voice should r  set only a small set of possible attributes; please feet  Definitions (scroll to see more)  Shift: A high pitched, pieroring, and sharp voice.  Basour, A round, neature, somewhat hash voice.  Basour, A round, neature, somewhat hash voice.  Paeker(P Please type out at least 3 distinct speech shyle	Information on the speaker's voice characteristics. In speaker's voice characteristics. In speaker's voice characteristics. In the primary speaker's voice. In the primary sp
<section-header><section-header><section-header><section-header><section-header><section-header><section-header><section-header><section-header><section-header><section-header><section-header><text><text><list-item><list-item><list-item><text><text><text><text><text><text><text><text><text><text></text></text></text></text></text></text></text></text></text></text></list-item></list-item></list-item></text></text></section-header></section-header></section-header></section-header></section-header></section-header></section-header></section-header></section-header></section-header></section-header></section-header>	speaker is provided. Please listen to the clip, paying carated by commas, that you think uniquely describe the ns and move on to the next audio clip.         see copy this code back to the Amazon Mechanical Turk each clips.         accords where you hear other speakers. Please focus or the basic characteristics of the speaker's voice should r         minimum difference of the speaker's voice should r         minimum difference of the speaker's voice should r         b basic characteristics of the speaker's voice should r         b basic characteristics of the speaker's voice should r         b basic characteristics of the speaker's voice should r         b basic characteristics of the speaker's voice should r         b basic characteristics of the speaker's voice should r         b basic characteristics of the speaker's voice should r         b basic characteristics of the speaker's voice should r         c basic characteristics of the speaker's voice should r         c basic characteristics of the speaker's voice should r         c basic characteristics of the speaker's voice should r         c Basic A white you the speaker and soothingly soft voice.         c Basic A nouch, pleasart and soothingly soft voice.         c Basic A nouch, aratine, somewhat hash voice.         pasker? Please type out at least 3 distinct speech style         c Basic A nouch speaker distingly couple for voice that conveys a speaker? Please type out at least 3 distinct speech style	Information to the speaker's voice characteristics. In speaker's voice characteristics. In the primary speaker's voice charac



1026		
1027		
1028		
1029		
1030		
1031		
1032		
1033		
1034		
1035		
1036	Instructions	•
1037	Welcome to our speech quality (naturalness and realisticity) evaluation task! Here are instructions on how	to use this interface:
1037	<ol> <li>Rate each clip jointly for how natural and realistic the speech sounds, on a scale of 1 (Bad) to 5 (Excellent).</li> <li>(Excellent) means the speech sounds very natural (e.g. spoken by a human) without robotic patterns and has</li> </ol>	1 (Bad) means that speech is low-quality, sounds very unnatural (e.g. robotic) or has low-quality audio and 5 high audio quality. Pay attention to only the naturalness and audio quality of the speech, not the content or
1030	the speaker's voice style. Sometimes, the audio may have partially uttered words at the beginning or the end 2. Note that the audio clips may have similar content, but each clip is different. Please rate each clip based or	; please ignore these. 1 how natural the speech sounds. You can compare the clips and rate them appropriately, giving similar
1039	ratings if you think the clips sound equally natural. 3. After selecting ratings, click the 'Save and Continue' button to move to the next annotation. Wait for the cli	ps to fully load.
1040	<ol> <li>Once you have completed all the audio clips, you will see a completion message with a survey code. Please</li> </ol>	e copy this code back to the Amazon Mechanical Turk task to receive your payment.
1041	<ol> <li>You can track how many examples you have annotated using the progress information right above the spece Note that you can collarse these instructions by clicking on the 'Instructions' text at the top.</li> </ol>	ech clips.
1042		
1043	R Clio1	
1044		Ate the quality (naturainess and realisticity) of the audio.     1: Bad     2: Poor     3: Fair     4: Good     5: Excellent
1045	000 All fine filler filler file and filler file and filler and filler and filler	
1046		
1047		
1048	22 Cup 2	Rate the quality (naturalness and realisticity) of the audio.
1049		A. bad 2. Pool 3. rail 4. 6000 3. Excettent
1050		
1051		
1052		Rate the quality (naturalness and realisticity) of the audio.
1053		1: Bad     2: Poor     3: Fair     4: Good     5: Excellent
1054	0:00 0:04	
1055		
1056	a Clip 4	Rate the quality (naturalness and realisticity) of the audio.
1057		1: Bad     2: Poor     3: Fair     4: Good     5: Excellent
1058	0:00 0:04	
1059		
1060	रः Clip 5	Rate the quality (naturalness and realisticity) of the audio.
1061		1: Bad     2: Poor     3: Fair     4: Good     5: Excellent
1062	0:00 0:05	
1063		
1064	@ Clip 6	Rate the quality (naturalness and realisticity) of the audio.
1065		1: Bad 2: Poor 3: Fair 4: Good 5: Excellent
1066	0:00 0:03	
1067		
1068	Caucand	Continue
1060	Save and	
1000		

Figure 6: Annotation UI for collecting Naturalness Mean Opinion Score ratings.



