Journal of Data-centric Machine Learning Research (2026) Submitted 12/25; Revised 09/26; Published 10/26

Vox-Profile: A Speech Foundation Model Benchmark for Characterizing Diverse Speaker Traits and Speech Properties

Tiantian Feng[1] TIANTIAF@USC.EDU
Jihwan Lee[1] JIHWAN@USC.EDU
Anfeng Xu[1] ANFENGXU@USC.EDU
Yoonjeong Lee[1] YOONJEOL@USC.EDU
Thanathai Lertpetchpun[1] LERTPETC@USC.EDU
Xuan Shi[1] XUANSHI@USC.EDU
Helin Wang[2] HWANG258@JHU.EDU
Thomas Thebaud[2] TTHEBAU1@JHU.EDU
Laureano Moro-Velazquez[2] LAUREANO@JHU.EDU
Dani Byrd[1] DBYRD@USC.EDU
Najim Dehak[2] NDEHAK3@JHU.EDU
Shrikanth Narayanan[1] SHRIKANN@USC.EDU

[1] *University of Southern California*
[2] *Johns Hopkins University*

Reviewed on OpenReview: *https://openreview.net/forum?id=qmGQoZfkEN*

Editor: Peter Mattson

Abstract

We introduce `Vox-Profile`, a comprehensive benchmark to characterize rich speaker traits and speech properties using speech foundation models. Unlike prior work that targets a single dimension of speaker traits, `Vox-Profile` provides holistic and multi-dimensional profiles spanning both static speaker traits (e.g., age, sex, accent, voice quality) and dynamic speech traits (e.g., emotion, speech flow). This benchmark is grounded in speech science and linguistics, developed with domain experts to accurately index speaker and speech characteristics. We report benchmark experiments using over 15 publicly available speech datasets and several widely used speech foundation models that target various static and dynamic speaker and speech properties. In addition to benchmark experiments, we showcase several downstream applications supported by `Vox-Profile`. First, we show that `Vox-Profile` can augment existing speech recognition datasets to analyze ASR performance variability. `Vox-Profile` is also used as a tool to evaluate the performance of speech generation systems. Finally, we assess the quality of our automated profiles through comparison with human evaluation and show convergent validity. `Vox-Profile` is publicly available at: https://github.com/tiantiaf0627/vox-profile-release.

Keywords: Speech Foundation Models, Speaker Traits, Accent Classification, Speech Emotion Recognition, Automatic Speech Recognition Analysis

©2026 Tiantian Feng, Jihwan Lee, Anfeng Xu, Yoonjeong Lee, Thanathai Lertpetchpun, Xuan Shi, Helin Wang, Thomas Thebaud, Laureano Moro-Velazquez, Dani Byrd, Najim Dehak, Shrikanth Narayanan.

1 Introduction

Even brief segments of voice carry a wealth of information about a speaker, including their age, sex, accent, emotion, and their physical environment and social context. The ability to accurately predict these attributes from speech has significant value in advancing a wide range of speech technologies. For example, conversational digital assistants like Alexa or Siri can adapt responses based on a user's perceived emotional state. Automatic speech recognition systems (ASR) (Prabhavalkar et al., 2023) can improve transcription accuracy by modeling and adapting to different accents. Despite these promising applications, much of the progress in speech modeling has been centered on conventional tasks such as ASR for transcribing spoken language, speaker diarization (Park et al., 2022) for tracking who is speaking when, and speech enhancement (Michelsanti et al., 2021) for improving audio speech quality. In contrast, there is limited work in systematically modeling and predicting diverse speaker and speech traits.

Much of the existing research characterizing speakers has focused on learning speaker embeddings, such as x-vectors (Snyder et al., 2018), which capture the unique vocal traits of individual speakers. These embeddings emerged as foundational components in many applications, most notably in speaker recognition and verification systems. While significant progress has been made in using speaker embeddings to capture identity-specific vocal features, relatively little work has focused on comprehensive modeling of other human-interpretable speaker traits, such as accent and voice quality (or timbre). Recent work has shown that robust inclusion of such traits holds great promise for downstream applications, like style-prompted speech generation models. For example, the `ParaSpeechCaps` (Diwan et al., 2025) dataset provides large-scale human annotations of speaker traits to facilitate the development of high-performing text-to-speech (TTS) models capable of generating speech with different speaking styles.

Although several studies have explored the modeling of speaker traits such as accents (Zuluaga-Gomez et al., 2023) or age (Burkhardt et al., 2023), each study designs its unique labeling choice in the classification, leading to inconsistencies across the literature. As a result, a unified framework for defining, categorizing, and modeling different speaker or speech traits is lacking. In this work, we propose `Vox-Profile`, one of the first benchmarking efforts that systematically evaluates rich multi-dimensional speaker and speech traits from English-speaking voices. Specifically, `Vox-Profile` experiments with over 15 publicly available datasets to predict static traits (e.g., age, sex, accent, and voice quality) and dynamic traits (e.g., speech emotion, speech flow) in different recording conditions and elicitation settings (e.g., read, spontaneous, and conversational speech). Our benchmark covers popular speech models, including `HuBERT` (Hsu et al., 2021a), `WavLM` (Chen et al., 2022), `ECAPA-TDNN` (Desplanques et al., 2020), and `Whisper` family (Radford et al., 2023).

Our proposed `Vox-Profile` benchmark exhibits multifold merits compared to prior works: (1) Unlike previous work on modeling limited speaker traits or speech properties, `Vox-Profile` provides a holistic characterization of an extensive list of traits from speech, including age, sex, accents, emotions, voice quality, speech flow, and expressiveness. (2) Instead of solely focusing on computational modeling, `Vox-Profile` integrates linguistically informed labeling choices to better address the inconsistent label choices of speaker traits across datasets. (3) The broad utility of `Vox-Profile` is established through investigating several potential

Table 1: Comparison of Vox-Profile with existing literature in modeling speaker and speech traits.

	Age	Sex	Accent	Categorized Emotion	Arousal Valence	Voice Quality	Speech Flow	Speech Expressiveness
VoxCeleb Enrichment (Hechmi et al., 2021)	✓	✓	✗	✗	✗	✗	✗	✗
CommonAccent (Zuluaga-Gomez et al., 2023)	✗	✗	✓	✗	✗	✗	✗	✗
GLOBE (Wang et al., 2024a)	✓	✓	✓	✗	✗	✗	✗	✗
ParaSpeechCaps (Diwan et al., 2025)	✗	✓	✓	✓	✗	✓	✗	✓
Vox-Profile (Ours)	✓	✓	✓	✓	✓	✓	✓	✓

applications, including analyzing speech model performance, evaluating speech generation systems, and automatically tagging speaking styles. Both machine and human evaluations indicate that `Vox-Profile` offers reliable estimations of speaker and speech traits.

2 Related Work

Static Trait Modeling. We summarize several related representative works in Table 1. Age and sex prediction are long-standing tasks in speaker trait modeling. Burkhardt et al.(Burkhardt et al., 2023) introduced a benchmark framework that uses several publicly available datasets for age and sex prediction based on the transformer. In a related effort, Yang et al.(Yang et al., 2025) investigated the effectiveness of `WavLM` embeddings for age and sex prediction, showing that self-supervised speech representation can achieve higher performances compared to i-vector features (Dehak et al., 2010). Apart from modeling age and sex attributes from speech, there has been growing interest in predicting speaker accents. For example, `CommonAccent` (Zuluaga-Gomez et al., 2023) introduced a benchmark for accent classification using samples from Mozilla Common Voice (Ardila et al., 2020). It reported strong performance in recognizing regional varieties of English, such as American, Canadian and British English, using the `ECAPA-TDNN` speaker embedding model. GLOBE (Wang et al., 2024a) is a similar effort that develops classifiers using `HuBERT` pre-trained models on Common Voice to predict accent as well as sex and age.

Dynamic Trait Modeling. Several works have also focused on modeling and characterizing dynamic speaker traits, such as speech emotion and fluency. One of the early efforts in supporting speech emotion recognition (SER) is the release of the IEMOCAP dataset (Busso et al., 2008), which provides annotated recordings of acted dialogues with fine-grained emotional labels to develop emotion recognition systems. Several datasets have since been designed to support more naturalistic emotion modeling, including the MSP-Podcast (Lotfian and Busso, 2017), MELD (Poria et al., 2019), CREMA-D (Cao et al., 2014), and CMU-MOSEI (Zadeh et al., 2018) datasets. Recent advances in speech foundation models have also led to the development of high-performing SER systems. Leveraging pre-trained representations from models such as `WavLM` and `Whisper`, researchers have demonstrated improvements over traditional SER approaches (Wagner et al., 2023; Feng and Narayanan, 2023). Likewise, a growing body of research focused on predicting speech fluency to develop robust language assessment systems. For example, SEP-28K (Lea et al., 2021) presents a notable effort to release large-scale speech data annotated for disfluencies, including conditions such as sound blocks.

Table 2: The labeling scheme of speaker traits and speech properties in `Vox-Profile` benchmark. The green, blue, and violet indicate speaker accents from North America, the British Isles, and other regions or language backgrounds, respectively. For voice quality, the colors green, blue, violet, olive, and orange represent dimensions of speaker pitch, rhythm, clarity, voice texture, and volume defined by `ParaSpeechCaps`. Finally, the blue color denotes labels related to speech disfluencies in the speech flow category.

Category	Traits	Labeling Scheme	Datasets
Static Traits	**Speaker Sex**	Male; Female	TIMIT (Garofolo et al., 1993); VoxCeleb (Nagrani et al., 2017); CommonVoice (Ardila et al., 2020)
	Speaker Age	Young adults (<30 Years); Adults (30-60 years); Senior adults (>60 years)	TIMIT (Garofolo et al., 1993); VoxCeleb (Nagrani et al., 2017); CommonVoice (Ardila et al., 2020)
	Speaker Accent	North America, English North America, Welsh, Scottish Northern Irish, Irish, Germanic Romance, Slavic, Semitic Oceania, South Africa, East Asia Southeast Asia, South Asia, Other	CommonVoice (Ardila et al., 2020); EdAcc (Sanabria et al., 2023); CSLU-FAE British Isles (Demirsahin et al., 2020b); TIMIT (Garofolo et al., 1993) L2-ARCTIC (Zhao et al., 2018); VoxPopuli (Wang et al., 2021); Fair-Speech (Veliche et al., 2024); ESLTTS (Wang et al., 2024b) Hispanic-English (Byrne et al., 2014); Nigerian-English (Ng, 2019)
	Voice Quality	Shrill, Nasal, Deep; Singsong, Pitchy, Flowing, Monotone, Staccato, Punctuated, Enunciated, Hesitant; Crisp, Slurred, Lisp, Stammering; Silky, Husky, Raspy, Guttural, Vocal-fry; Booming, Authoritative, Loud, Hushed, Soft	ParaSpeechCaps (Diwan et al., 2025)
Dynamic Traits	**Categorical Emotion**	Neutral, Happy, Sad, Angry, Contempt Fear, Disgust, Surprise, Other	MSP-Podcast (Lotfian and Busso, 2017)
	Arousal/Valence	0-1(Calm->Active) / 0-1(Negative->Positive)	MSP-Podcast (Lotfian and Busso, 2017)
	Speech Flow	Fluent, Disfluent (Prolongation, Word Repetition Sound Repetition, Block, Interjection)	Sep-28K (Lea et al., 2021) FluencyBank (Ratner and MacWhinney, 2018)
	Speech Expressiveness	Animated, Laughing, Passive Whispered, Enunciated	ParaSpeechCaps (Diwan et al., 2025)

3 Labeling Choice of Speaker Traits and Speech Properties

Unlike prior benchmarks focused on general speech modeling tasks (Yang et al., 2021), `Vox-Profile` offers a structured benchmarking framework for systematically predicting a broad range of speaker traits and speech properties. In developing this automated evaluation pipeline, we observed considerable inconsistency across existing studies in how speaker traits and speech properties are categorized, defined, and modeled. One of the closest prior efforts is `ParaSpeechCaps` (Diwan et al., 2025), which proposes a labeling choice that models "intrinsic" (e.g., accent) and "situational" (e.g., emotion) traits. However, its categorization remains underspecified and difficult to reconcile with conventions based on empirical foundations in speech science literature. For example, while `ParaSpeechCaps` includes a broad set of emotional descriptors, these labels deviate from standard categorical and dimensional emotion frameworks commonly used in speech and language research, creating challenges when aligning datasets or comparing across studies.

Many existing approaches in speech trait modeling lack grounding in speech science and linguistics, which often leads to ambiguous or ill-posed problem formulations. A prominent example is speaker age modeling, which is frequently treated as a regression task targeting a speaker's exact chronological age (Si et al., 2022; Alharthi et al., 2025). However, this framing overlooks numerous confounding variables such as biological sex, vocal anatomy, and speaking habits that affect vocal aging. Crucially, it misaligns with how age is typically perceived (and experienced) from speech: human listeners estimate age in broad *intervals* rather than in precise years (Schötz, 2006; Skoog Waller et al., 2015). Similarly, accent

classification is often based on a speaker's nationality, as in `ParaSpeechCaps`, even though nationality is not reliably inferable from speech alone, particularly in multilingual or diasporic contexts (Rubin, 1992; Humayun et al., 2024). To address these challenges, we introduce a linguistically principled labeling choice in `Vox-Profile`, designed to be generalizable across diverse speech datasets and conditions.

We define two broad categories of speech traits to structure the benchmark: (1) Static speaker traits capture relatively stable characteristics of the speaker: age, sex, accent, and voice quality; and (2) dynamic speech traits reflect context-dependent aspects of speech that can vary across situations, including emotion, speech expressiveness, and speech flow. The full labeling choices are summarized in Table 2, with the description of all datasets used in this benchmark and data splits for the benchmark experiment detailed in Appendix A.

3.1 Static Speaker Traits

Speaker Sex Biological sex as reflected in voice is relatively well-defined in the literature. We follow standard practice by modeling it as a binary classification (male vs. female), while recognizing that, in practice, acoustic parameters of voice quality do not always map cleanly onto binary categories, particularly in the case of young children or speakers with atypical vocal profiles.

Speaker Age As discussed earlier, many existing studies treat age estimation as a regression task, aiming to predict a speaker's exact chronological age from speech (Si et al., 2022; Alharthi et al., 2025). We argue that this formulation is ill-posed: perceived vocal age is shaped by multiple interacting factors and is more naturally categorized in intervals than estimated precisely. Instead, we propose classifying speakers into three broad age groups: young adults (18-30 years), adults (30-60 years), and senior adults (>60 years). Speech science research (Reubold et al., 2010) indicates that the vocal characteristics typically stabilize in early adulthood (approximately 18 years), with relatively consistent voice characteristics across the adult decades. In contrast, age-related vocal changes, linked to hormonal, physiological, and health-related factors, become increasingly salient from around age 60 years onward.

Speaker Accent An accent refers to a distinctive way of pronouncing a language, typically shaped by a speaker's regional origin and linguistic background. However, the classification of accents has varied widely across the literature with inconsistent labeling schemes and taxonomic assumptions (Wang et al., 2024a; Zuluaga-Gomez et al., 2023; Zhong et al., 2025). In this work, we introduce a unified and scalable labeling choice for accent classification, designed to integrate multiple mainstream English-accent datasets under a coherent framework. Focusing on English-speaking populations, we first organize accents into three broad regional groups: North American, British Isles, and other regions or other language backgrounds (e.g., Oceania, South Asia, and Africa). Within the British Isles, we further distinguish specific regional varieties: English (England), Scottish, Northern Irish, Welsh, and Irish. To better capture systematic cross-linguistic influences on English pronunciation, we also group certain English accents by the language family of the speakers' first language (L1). This includes Germanic (e.g., German, Dutch), Slavic (e.g., Russian, Polish), Romance (e.g., Spanish, French, Italian), and Semitic (e.g., Arabic, Hebrew) language backgrounds. This typological

grouping reflects the fact that related L1s often give rise to similar segmental and prosodic transfer patterns in second-language (L2) English, which do not always align clearly with national or geographic borders. Asian accents are organized by region, East Asia, South Asia, and Southeast Asia, reflecting common typological and areal features. Additional categories include Oceania and South Africa, representing other major English-speaking regions. Due to underrepresentation in existing datasets, some language backgrounds are not able to be classified into these categories reliably. For example, accents associated with Uralic languages as L1s (e.g., Finnish) or with regions of Africa outside South Africa are grouped under the general "Other" label. This labeling choice enables the integration of 11 publicly available datasets, each with distinct accent labeling conventions, into a unified benchmark for training and evaluating accent classification models.

Voice Quality Only a few publicly available datasets offer detailed annotations suitable for modeling voice quality. For this benchmark, we adopt `ParaSpeechCaps` (Diwan et al., 2025), the dataset with the most extensive human-annotated labels currently available in this category. It defines voice quality across five perceptual dimensions: pitch, voice texture, volume, clarity, and rhythm, each reflecting a distinct aspect of how a speaker's voice is perceived beyond segmental content. The specific labels associated with each dimension are provided in Table 2.

3.2 Dynamic Speech Traits

Categorical Emotion There is an extensive list of SER datasets, including IEMOCAP (Busso et al., 2008), MELD (Poria et al., 2019), CMU-MOSEI (Zadeh et al., 2018), and CREMA-D (Cao et al., 2014), each employing distinct emotion labeling schemes. Among them, we find the MSP-Podcast dataset (Lotfian and Busso, 2017) provides the largest number of speech emotion samples recorded in naturalistic settings. Therefore, we adopt its emotion labeling scheme, including categories neutral, happy, sad, angry, fear, disgust, contempt, surprise, non-agreement, and others.

Arousal/Valence Arousal and valence are also widely used to describe emotional states in speech (Kuppens et al., 2013). We define both arousal and valence as continuous variables scaled from 0 to 1. *Arousal* represents the intensity of the emotion, with higher scores indicating greater activation or energy. *Valence* reflects emotional polarity, with lower scores corresponding to more negative affective states (e.g., sadness) and higher values to more positive (e.g., happiness).

Speech Flow Speech flow has gained growing attention in recent years as a dimension of fluency in spoken language (Ratner and MacWhinney, 2018). One of the most notable resources is the Sep-28K dataset (Lea et al., 2021), which annotates disfluencies at scale. Following this foundation, we categorize speech flow into two: fluent and disfluent. Disfluency is further defined to include prolongations, word repetitions, sound repetitions, blocks, and interjections (including filler words or hesitation sounds).

Speech Expressiveness The last dynamic trait we aim to model is speech expressiveness. Similar to voice quality, the datasets available are scarce in this field. We adopt the definitions provided in the `ParaSpeechCaps` (Diwan et al., 2025) dataset to guide our modeling.

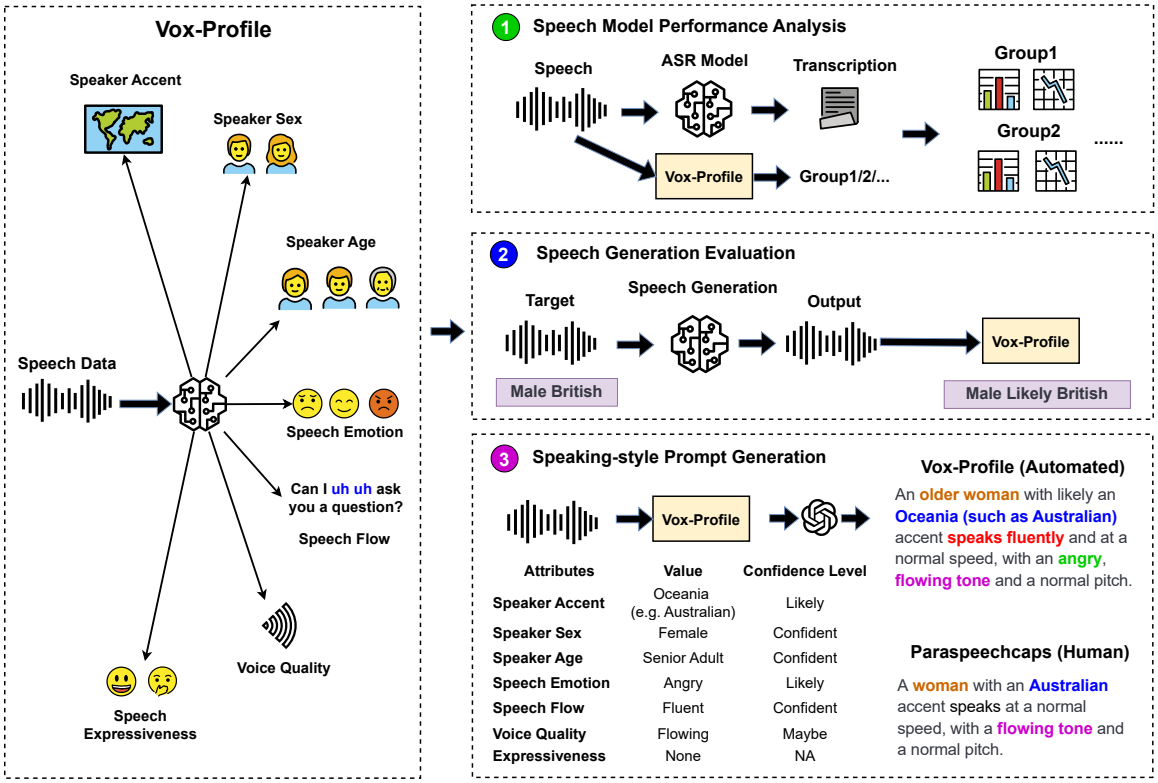


Figure 1: Overview of the `Vox-Profile` Benchmark and its applications. We highlight three primary use cases: (1) speech model (such as ASR) performance analysis, (2) automated speech generation evaluation, and (3) automated speaking style tagging. The figure uses sources from the OpenMoji web.

Specifically, we categorize speech expressiveness into five types: animated, laughing, passive, whispered, and enunciated.

4 Vox-Profile Benchmark

As shown in Figure 1, we propose a speaker traits and speech property modeling benchmark, `Vox-Profile`, to model a wide range of speaker traits and speech properties defined above. Our `Vox-Profile` benchmark can be applied to augment metadata for analyzing speech model performances, evaluating speech generation systems, and curating synthetic speaking style descriptions.

4.1 Benchmark Overview

Static Speaker Traits We examine the aforementioned speech foundation models to predict different static traits from speech. (1) For age and sex classification, we experiment with the multitask learning approach to jointly predict both traits. This is motivated by the existing studies that show changes in voice characteristics across age can differ significantly between males and females. Modeling them jointly is expected to allow the system to better

capture these interactions and improve classification performance. We apply the concordance correlation coefficient (CCC) (Lawrence and Lin, 1989) loss to train age prediction systems. Instead of reporting regression results from absolute age values, we map ages into the three predefined age groups and report the corresponding classification performance. (2) We develop accent classification systems in two levels of categorization: broad and narrow. The former targets broader accent categories, while the latter predicts 16 narrow accent labels based on different regional and language backgrounds. (3) We formulate voice quality prediction as a multi-label classification task since each speech sample can be associated with multiple voice qualities.

Dynamic Speech Traits (1) Our categorical emotion and dimensional emotion (arousal / valence) modeling systems are built upon our recent top-ranked solutions (Lertpetchpun et al., 2025; Feng et al., 2025) in the Speech Emotion Recognition in Naturalistic Conditions Challenge at Interspeech 2025 (IS25-SER challenge) [1]. Specifically, the categorical emotion system applies the KL divergence loss to predict the distribution of categorical emotion annotations. For arousal and valence modeling, we use the sigmoid function to map arousal and valence values onto a normalized scale between 0 and 1. (2) For speech flow modeling, we propose a multitask learning framework of 2 classification heads with one to predict whether the speech is fluent or disfluent, the other to classify the specific types of disfluencies present. We highlight that classifying types of disfluencies is a multilabel task, as multiple breaks can co-occur in a single speech utterance. (3) We formulate speech expressiveness as a multilabel classification task.

4.2 Data Preprocessing

For all experimental datasets used in speaker traits and speech properties classification, we resample the audio to 16 kHz. We exclude audio samples shorter than 3 seconds, as reliably estimating speaker traits or speech properties is not feasible with shorter speech utterances. We also truncate the audio to 15 seconds due to the computation limitation in training. Audios with corrupted formats are removed from the experiments. For speaker accent classification, we discard samples with labels like British, as it does not specify the regional varieties.

4.3 Modeling Approach

In `Vox-Profile`, we evaluate several widely studied speech foundation models, including the `HuBERT` (Hsu et al., 2021a), `WavLM` (Hu et al., 2024), `ECAPA-TDNN` (Desplanques et al., 2020), and `Whisper` (Radford et al., 2023) family with details reported in Appendix B. Since speaker embedding models are primarily related to static speaker traits, we exclude the `ECAPA-TDNN` model for predicting dynamic speech traits. For all experiments, we use the downstream model architecture developed in (Pepino et al., 2021), where many existing works have shown that this simple architecture achieves strong performance in a wide range of downstream speech modeling tasks. Specifically, we first apply weighted averaging to combine hidden representations of both the convolutional and Transformer (Vaswani et al., 2017) encoder layers. The aggregated output is then processed through 1D-pointwise convolutional layers.

1. https://lab-msp.com/MSP-Podcast_Competition/IS2025

Table 3: Comparison of different speech foundation models in predicting static traits. Overall, the results show that `Whisper` Large achieves the overall best performance, while the `WavLM`-based model shows moderate strength in accent prediction.

	Speaker Age		**Speaker Sex**		**Broad Accent**		**Narrow Accent**		**Voice Quality**
	Acc	**F1**	**Acc**	**F1**	**Acc**	**F1**	**Acc**	**F1**	**F1**
Speaker Model									
↳ECAPA-TDNN	59.3	0.469	96.9	0.962	74.6	0.638	42.3	0.363	0.706
Self-Supervised									
↳HuBERT Large	57.1	0.501	96.7	0.960	85.4	0.841	49.1	0.443	0.692
↳WavLM Large	67.6	0.624	97.7	0.971	**92.9**	0.884	65.7	0.617	0.703
Whisper Family									
↳Whisper Tiny	58.1	0.520	94.3	0.934	89.1	0.831	54.1	0.484	0.587
↳Whisper Small	65.8	0.558	96.4	0.958	90.9	0.863	70.4	0.663	**0.737**
↳Whisper Large	**69.5**	**0.643**	**98.0**	**0.975**	92.8	**0.889**	**75.8**	**0.724**	0.704

Finally, we average the convolutional outputs to obtain the final embeddings, which are passed through fully connected layers for classification or regression. In addition to this approach, we experiment with finetuning using Low-Rank Adaptation, `LoRa` (Hu et al., 2022), an effective method for adapting speech foundation models. Details of the training hyperparameters, such as learning rates and training epochs, and training resources, are provided in Appendix C and D.

4.4 Evaluation Metrics

We compare different models in predicting speaker and speech traits, including biological sex, age, accent, categorical emotions, and speech fluency/disfluency, using both accuracy and macro F1 scores. Additionally, we report the macro F1 score when comparing models predicting speech voice quality, specific speech disfluencies, and speech expressiveness. Finally, we assess the predictions of arousal and valence using the concordance correlation coefficient (CCC) (Lawrence and Lin, 1989) score.

5 Benchmark Performance

In this section, we present a detailed comparison of benchmark performance. We report the top-performing scores for each model in Table 3 and Table 4. We then analyze the impact of model ensembling, followed by a comparison with existing works from the literature.

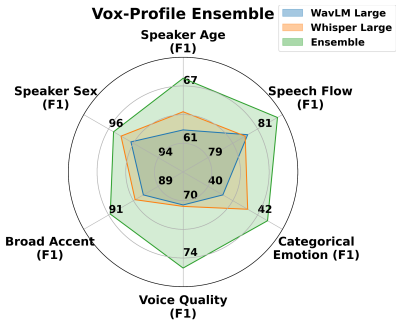
5.1 Overall Benchmark Performance

Static Speaker Traits We compare the performance of different speech foundation models in predicting static speaker traits shown in Table 3. The results indicate a positive correlation between model parameter size and performance, with larger models generally achieving better performance in speaker trait classification. Moreover, although the `ECAPA-TDNN` model is optimized for speaker recognition tasks, finetuning `ECAPA-TDNN` still yields lower performance than self-supervised or supervised speech foundation models, particularly in accent and age classification. We observe that sex classification is relatively straightforward, with most models achieving around 95% accuracy. Moreover, classifying between broad accent groups

Table 4: Comparison of different models in predicting dynamic speech traits.

	Speech Emotion F1	Arousal CCC	Valence CCC	Fluency Acc	Fluency F1	Disfluency Type F1	Speech Expressivenss F1
Self-Supervised							
↳HuBERT Large	0.395	0.585	0.650	75.5	0.755	0.627	0.565
↳WavLM Large	0.406	0.585	**0.665**	**80.7**	**0.806**	**0.691**	0.689
Whisper Family							
↳Whisper Tiny	0.337	0.545	0.522	76.4	0.762	0.639	0.484
↳Whisper Small	0.387	0.583	0.623	80.4	0.804	0.653	0.751
↳Whisper Large	**0.416**	**0.588**	0.645	80.6	0.805	0.670	**0.848**

Table 5: Experiments to compare `Vox-Profile` with existing work and impact of model ensembles.



(a) Comparing model ensembles with single model in `Vox-Profile`.

Model	Acc	F1	Model	Acc	F1
Accent (VCTK)			**Accent (British Isles)**		
↳CommonAccent	66.3	0.379	↳CommonAccent	69.4	0.581
↳Vox-Profile	**72.6**	**0.423**	↳Vox-Profile	**83.9**	**0.770**
Speaker Age (VoxCeleb)			**Speaker Sex (TIMIT)**		
↳Wav2Vec2-based	74.9	**0.717**	↳Wav2Vec2-based	99.4	0.993
↳Vox-Profile	**75.6**	0.716	↳Vox-Profile	**99.5**	**0.995**
Model	**F1**		**Model**	**CCC**	
Categorical Emotion			**Arousal / Valence**		
↳IS25-SAIL SER	**0.428**		↳IS25-SAIL SER	**0.642** / **0.683**	
↳Vox-Profile	0.392		↳Vox-Profile	0.623 / 0.649	

(b) Comparing `Vox-Profile` benchmark performance with the existing literature.

(North American, British Isles, and other regional or language backgrounds) yields strong performance, with macro F1 scores exceeding 0.80 in most models. However, narrow accent prediction remains challenging with only `Whisper Large` achieving a 0.724 macro F1 score in this task.

Dynamic Speech Traits We present results of dynamic traits prediction in Table 4. Consistent with findings for static trait prediction, we find that larger models perform better, with `Whisper Large` and `WavLM Large` achieving the highest scores across most tasks. Moreover, detecting disfluency in speech flow is relatively straightforward, with multiple models achieving over 80% accuracy. In contrast, the benchmark results indicate that SER remains challenging given the relatively low performance across most experiments. We highlight that most models can only reach macro F1 scores around 0.4 (including our models ranked as the top-performing solutions in the IS25-SER challenge).

Model Ensemble We investigate whether ensembling top-performing models further improves the performance. The results in Table 5a show that naïve ensembles of the two top-performing models consistently yield moderate improvements compared to individual models. In particular, we observe substantial gains in modeling static traits such as speech voice quality, age, and sex.

5.2 Comparing Vox-Profile Benchmark with Existing Literature

We compare the benchmark performance of `Vox-Profile` to several prior works. Regarding the accent classification, we compare our model against a previous approach CommonAccent (Zuluaga-Gomez et al., 2023) on the VCTK (Veaux et al., 2012) and British Isles (Demirsahin et al., 2020a) speaker datasets. To ensure a fair comparison, we mapped the label schema used in CommonAccent to align with our accent labeling. As shown in Table 5b, our benchmark model consistently outperforms CommonAccent in both datasets. Moreover, we compare our speaker age and sex classification model with the approach presented in (Burkhardt et al., 2023), using the test sets of the VoxCeleb and TIMIT datasets, respectively. The results indicate that `Vox-Profile` achieves strong performance in both age and sex prediction tasks. Finally, we compare our SER models with our top-performing system, `SAIL-SER`, which achieves top-1 and top-2 scores in the categorical and dimensional SER in the IS25-SER challenge. Note that `SAIL-SER` utilizes both text and speech modalities, while `Vox-Profile` relies solely on the speech input. The absence of text modality in `Vox-Profile` explains the performance gap; however, it still achieves comparable performance despite this limitation.

6 Enabling Versatile Speech Applications with Vox-Profile

6.1 Speech Model Performance Analysis

We show how `Vox-Profile` facilitates the analysis of speech model performance. Specifically, we generate speaker and speech traits for existing datasets and investigate whether these generated labels can lead to the same insights as using the ground truth trait information in analyzing the speech model performances. For this experiment, we focus on ASR as a representative task in speech modeling.

ASR Inference We augment accent and emotion labels for the VCTK and MSP-Podcast datasets. For MSP-Podcast, we provide emotion labels only for the dev set, following the 4 most frequently represented emotions: neutral, sadness, happiness, and anger. Only samples with confident predictions are used for computing ASR performance in the `Vox-Profile` experiment. To mitigate the potential artifacts from using the same model architecture for ASR and speech trait prediction, we use the `Wav2Vec 2.0 Robust` (Hsu et al., 2021b) to infer the transcript. In our analysis, we exclude the ASR models, which the VCTK dataset is trained on, such as `Parakeet` (Rekesh et al., 2023) or `Canary` (Rekesh et al., 2023), for fair comparison.

Findings Figure 2 shows that ASR performance trends based on the ground truth trait align closely with those using the predicted labels from `Vox-Profile`. In the VCTK dataset, North American speakers consistently have lower WER scores than speakers from the British Isles and those with "Other" accents, regardless of whether ground truth or synthetic labels are used ($p < 0.01$ in both conditions). Moreover, whether using ground truth or predicted emotion labels, in this particular dataset, speech expressing sadness is associated with relatively lower WER scores, whereas happy speech tends to result in higher WER scores. Overall, these findings suggest that `Vox-Profile` can generate reliable synthetic metadata to facilitate in-depth analysis of speech model performance.

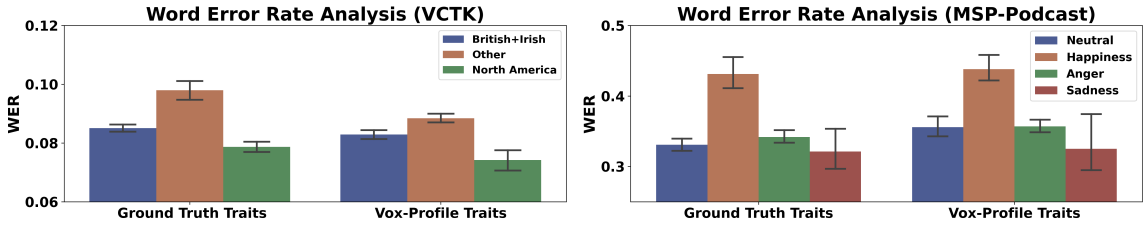


Figure 2: ASR performance trends, grouped by ground truth and predicted labels by `Vox-Profile`. We measure WER, stratified by accent and emotion labels. We observe similar performance trends between the predicted and ground truth trait labels.

6.2 Automated Evaluation Tool for Speech Generation Tasks

We demonstrate the utility of `Vox-Profile` as an evaluation tool for speech generation tasks by comparing two representative models: `FreeVC` (Li et al., 2023) and `VALLE-X` (Zhang et al., 2023). `FreeVC` is selected as a representative of textless voice conversion models that operate in the latent space, aiming to transform source speech to match the voice identity of a reference speaker. `VALLE-X` is chosen as a representative of voice cloning approaches based on neural codecs, utilizing a concatenated pipeline of ASR and TTS conditioned on the reference speech. It has been reported that textless models like `FreeVC` often struggle to accurately reflect the accent of the reference speaker (Baade et al., 2024). On the other hand, `VALLE-X`, by incorporating an intermediate text representation through ASR, is better equipped to preserve accentual features in the generated speech.

To assess the ability of these models to reflect the accent of the reference speech, we randomly selected source-reference (target) pairs from the VCTK dataset (Veaux et al., 2012). For each pair, we synthesized speech samples and evaluated whether they more closely resembled the source or reference speaker in terms of accent, by measuring cosine similarities and accent prediction scores. Detailed configurations for `FreeVC`, `VALLE-X`, and the selected test speakers are provided in Appendix D.1.

As shown in Table 6, the accent prediction scores and the cosine similarity for the synthesized samples from `FreeVC` suggest greater similarity to the source speaker's accent than to the reference speaker. In contrast, the scores for `VALLE-X` indicate closer alignment with the reference speaker's accent in most conditions. These findings are consistent with previous studies, which report that `FreeVC` has limited capability in replicating the accentual features of the reference speaker, whereas `VALLE-X`, due to its intermediate text-based representation, more effectively preserves these features (Baade et al., 2024; Zhang et al., 2023).

6.3 Generating Synthetic Speaking Style Prompt

We further apply the `Vox-Profile` as a tool for generating synthetic speaker and speech traits to create speaking style prompts. Previous efforts to create synthetic speaking style prompts, such as `ParlerTTS` (Lyth and King, 2024), rely on heuristic-based categorizations of acoustic features, including fundamental frequency (F0) and signal-to-noise ratio (SNR). More recently, `ParaSpeechCaps` introduced a large-scale, human-annotated dataset containing

Table 6: Utilizing `Vox-Profile` to evaluate the accent conversion performance of `FreeVC` and `VALLE-X`, by measuring cosine similarity (left table) and prediction scores (right table).

Source ⇒ Reference	Similarity Between (Source, Output) ↓	(Reference, Output) ↑
British Isles ⇒ North America		
↳FreeVC	**0.874 ± 0.046**	0.319 ± 0.071
↳VALLE-X	0.512 ± 0.103	**0.543 ± 0.080**
British Isles ⇒ Other		
↳FreeVC	**0.818 ± 0.064**	0.552 ± 0.089
↳VALLE-X	0.497 ± 0.102	**0.908 ± 0.055**
North America ⇒ British Isles		
↳FreeVC	**0.846 ± 0.046**	0.503 ± 0.086
↳VALLE-X	0.559 ± 0.082	**0.595 ± 0.090**
Other ⇒ British Isles		
↳FreeVC	0.937 ± 0.036	**0.658 ± 0.090**
↳VALLE-X	**0.960 ± 0.030**	0.552 ± 0.097

British Isles -> North America			
Prediction Scores	**British Isles**	**North America**	**Other**
Source	**55.1**	0.8	44.1
Reference	7.5	**57.6**	35.0
FreeVC Output	**59.8**	0.8	39.5
VALLE-X Output	1.9	11.1	**86.9**
British Isles -> Other			
Prediction Scores	**British Isles**	**North America**	**Other**
Source	**55.1**	0.8	44.1
Reference	5.3	2.3	**92.4**
FreeVC Output	**56.7**	0.8	42.5
VALLE-X Output	2.5	11.2	**86.3**

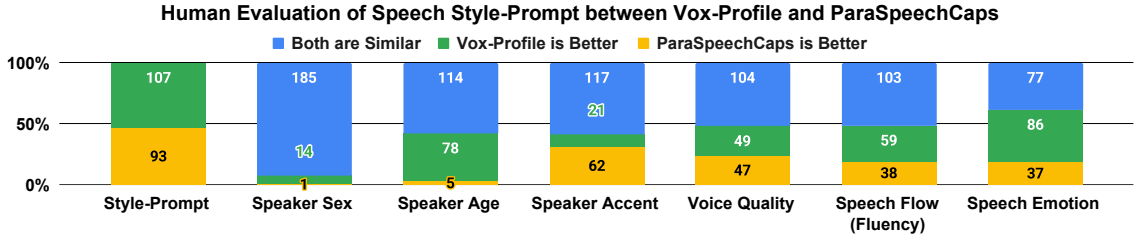


Figure 3: Comparing human evaluation results between synthetic speaking style prompts created by `Vox-Profile` and human-annotated prompts from `ParaSpeechCaps`. Human raters provided their preferences across overall prompt quality, sex, age, accent, voice quality, fluency, and emotion.

multiple speech traits, such as voice quality and speaker accent. In contrast to these prior works, `Vox-Profile` provides a more extensive and varied set of traits, including speech flow, arousal, valence, and speaker age. Moreover, computational models of `Vox-Profile` output probabilistic predictions for each trait, enabling more nuanced and confidence-sensitive descriptions. For example, a Scottish accent prediction with a probability of 0.9 can be described as having a distinct Scottish accent, while a probability of 0.5 might be described as a likely Scottish accent. This probabilistic framing supports the generation of speech descriptions that better reflect the para-linguistic uncertainties with which humans naturally perceive and interpret speech for information extraneous to linguistic messaging. To evaluate the robustness and versatility of the style prompts generated by `Vox-Profile`, we propose a comparative analysis between human-annotated and machine-generated style prompts using the `ParaSpeechCaps` dataset. The detailed human evaluation procedure is described in Appendix F.

Prompt Generation Since our voice quality and expressiveness models are trained on the training set of the `ParaSpeechCaps`, we use speech samples from the test set for this experiment. Specifically, we select 30 speech samples stratified by speaker sex and accent (broader accent classes). For each selected sample, we use `Vox-Profile` to infer both static and dynamic traits. To ensure reliability, we keep the predictions with high and

moderate confidence levels, labeling them as "distinct/very likely" and "likely," respectively. Traits with low confidence are excluded from prompt generation. Subsequently, we use `GPT-4.1` (OpenAI, 2024) to generate the speaking style prompt using the synthetic tags. To ensure a fair comparison, we also apply the `GPT-4.1` to generate the speaking style prompt using the `ParaSpeechCaps` tags. The prompt generation follows the template used in `ParaSpeechCaps` and the detailed description of confidence levels and prompt designs is in Appendix E.

Human-Evaluation Results Human-evaluation results comparing synthetic speaking style prompts from `Vox-Profile` and human-annotated speaking style prompts from ParaSpeechCaps are presented in Figure 3. Overall, the results suggest that this group of human raters shows similar preference levels for both synthetic and human-annotated speaking style prompts. Specifically, they favor the emotion, age, and speech flow descriptions generated by `Vox-Profile` over those from `ParaSpeechCaps`. For accent descriptions, the human raters reach similar levels of agreement between `ParaSpeechCaps` and `Vox-Profile` in more than 50% unique evaluations. However, ground truth accent labels from `ParaSpeechCaps` still outperform `Vox-Profile` in accurately describing speaker accents, highlighting that accent prediction remains challenging in speaker trait modeling. In summary, the human evaluation results provide evidence suggesting that `Vox-Profile` can be used to create synthetic speaking style prompts that closely match human-labeled data.

7 Ethical Use and Mitigation of Misuses

It is critical to recognize that `Vox-Profile` predicts speaker group-level traits and does not perform speaker recognition. Privacy risks are therefore different but not absent. The intended use of `Vox-Profile` includes analysis, evaluation, and dataset augmentation, and it should not be used for individual profiling or eligibility decisions. While speaker recognition is used to uniquely identify individuals from speech, raising direct concerns about surveillance or re-identification, Vox-Profile instead infers generalized traits such as age group, which are less personally identifiable. Nonetheless, we acknowledge that under different conditions, even these trait predictions may pose privacy risks depending on legal definitions, usage agreements, or industry practices. To mitigate potential misuses, we discuss two complementary prevention strategies below.

Legal Prevention Most `Vox-Profile` models are released under a Responsible AI License (RAIL) that explicitly prohibits applications involving surveillance, discrimination, or unauthorized inference of sensitive personal data under the scope of law and regulation. For example, the license prohibits using the system to "detect or infer any legally protected class or aspect of any person" or to "diagnose a medical condition without human oversight" Users should respect the privacy and consent of the data subjects and adhere to the relevant laws and regulations in their jurisdictions when using `Vox-Profile`. Specifically, models derived from data hosted using FluencyBank (Ratner and MacWhinney, 2018), part of the TalkBank system (MacWhinney), are subject to TalkBank's Ground Rules and distributed under CC BY-NC-SA 3.0. Moreover, we encourage users to apply `Vox-Profile` only for robustness analysis and evaluation of speech technologies and not for individual re-identification. The allowed applications include evaluating limitations in ASR models or improving speech

generation systems, as shown in the paper. Relevant ethical considerations include, but are not limited to, those outlined by Gao et al. (Gao et al., 2024).

Software Users may consider voice anonymization techniques before applying `Vox-Profile` to prevent revealing sensitive data. For example, the researchers may consider adding noise to specific features, as shown by (Mireshghallah et al., 2021), to anonymize speaker age, but not speaker accent, when evaluating ASR models for accented voices, when consent allows such use. Users may refer to more voice anonymization techniques described in (Tomashenko et al., 2024).

8 Conclusion and Limitations

We propose `Vox-Profile`, a comprehensive benchmark for modeling speaker and speech characteristics. This benchmark includes large-scale experiments across more than 15 public speech datasets, evaluated using widely adopted speech foundation models, and delivers a suite of high-performing trait prediction models. Our results show that `Vox-Profile` supports a broad range of applications, including speech model performance analysis, generated speech evaluation, and high-quality speaker and style prompt generation. Our next steps include testing our benchmark in multilingual conditions and adding the prediction of multilingual speech attributes like language identification and code-switching to improve accessibility for more diverse language communities. Moreover, while `Vox-Profile` currently uses speech models pre-trained mostly for ASR, we plan to expand to alternative architectures such as `Emotion2Vec` (Ma et al., 2024) and `Speech-LLMs` (Fathullah et al., 2024; Hu et al., 2024), which we expect will broaden the benchmark's applicability to affective and conversational speech technologies.

Limitations Firstly, the human evaluation in Section 6.3 is a small-scale study involving approximately 10 raters and 30 speech samples. Because the raters consist primarily of male, non-native English speakers (as shown in Appendix F), the current findings may not generalize to broader listener populations. Specifically, a larger study with a more demographically balanced group of raters is needed to evaluate whether perceptions of prompt quality differ across listener groups. One other limitation is that we have not systematically verified whether our evaluation datasets overlap with the data used to pre-train the speech foundation models. For example, some Vox-Profile datasets, including Common Voice and VoxCeleb, may appear in the published pretraining or fine-tuning corpora of related speech models. Specifically, we cannot exclude potential overlap with the data used for the Whisper checkpoints.

9 Acknowledgement

This work was supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the ARTS Program under contract D2023-2308110001. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

Crowdsourced high-quality nigerian english speech data set. *Open Speech and Language Resources*, 2019.

Dareen Alharthi, Mahsa Zamani, Bhiksha Raj, and Rita Singh. Tessellated linear model for age prediction from voice. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–4, 2025. doi: 10.1109/ICASSP49660.2025.10888079.

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

Alan Baade, Puyuan Peng, and David Harwath. Neural codec language models for disentangled and textless voice conversion. In *Interspeech 2024*, pages 182–186, 2024. doi: 10.21437/Interspeech.2024-1298.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

Felix Burkhardt, Johannes Wagner, Hagen Wierstorf, Florian Eyben, and Björn Schuller. Speech-based age and gender prediction with transformers. In *Speech Communication; 15th ITG Conference*, pages 46–50. VDE, 2023.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359, 2008.

William Byrne, Eva Knodt, Jared Bernstein, and Farzhad Emami. Hispanic-english database (ldc2014s05). *Linguistic Data Consortium*, 2014.

Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390, 2014.

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

Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet. Front-end factor analysis for speaker verification. *IEEE Transactions on Audio, Speech, and Language Processing*, 19(4):788–798, 2010.

Isin Demirsahin, Oddur Kjartansson, Alexander Gutkin, and Clara Rivera. Open-source Multi-speaker Corpora of the English Accents in the British Isles. In *Proceedings of The 12th Language Resources and Evaluation Conference (LREC)*, pages 6532–6541, Marseille, France, May 2020a. European Language Resources Association (ELRA). ISBN 979-10-95546-34-4. URL https://www.aclweb.org/anthology/2020.lrec-1.804.

Isin Demirsahin, Oddur Kjartansson, Alexander Gutkin, and Clara Rivera. Open-source multi-speaker corpora of the english accents in the british isles. In *Proceedings of the twelfth language resources and evaluation conference*, pages 6532–6541, 2020b.

Brecht Desplanques, Jenthe Thienpondt, and Kris Demuynck. Ecapa-tdnn: Emphasized channel attention, propagation and aggregation in tdnn based speaker verification. In *21st Annual conference of the International Speech Communication Association (INTERSPEECH 2020)*, pages 3830–3834. International Speech Communication Association (ISCA), 2020.

Anuj Diwan, Zhisheng Zheng, David Harwath, and Eunsol Choi. Scaling rich style-prompted text-to-speech datasets. *arXiv preprint arXiv:2503.04713*, 2025.

Yassir Fathullah, Chunyang Wu, Egor Lakomkin, Ke Li, Junteng Jia, Yuan Shangguan, Jay Mahadeokar, Ozlem Kalinli, Christian Fuegen, and Mike Seltzer. Audiochatllama: Towards general-purpose speech abilities for llms. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5522–5532, 2024.

Tiantian Feng and Shrikanth Narayanan. Peft-ser: On the use of parameter efficient transfer learning approaches for speech emotion recognition using pre-trained speech models. In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 1–8. IEEE, 2023.

Tiantian Feng, Thanathai Lertpetchpun, Dani Byrd, and Shrikanth Narayanan. Developing a top-tier framework in naturalistic conditions challenge for categorized emotion prediction: From speech foundation models and learning objective to data augmentation and engineering choices. In *Interspeech2025*, 2025.

Haoyu Gao, Mansooreh Zahedi, Christoph Treude, Sarita Rosenstock, and Marc Cheong. Documenting ethical considerations in open source ai models. In *Proceedings of the 18th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, pages 177–188, 2024.

John S Garofolo, Lori F Lamel, William M Fisher, Jonathan G Fiscus, and David S Pallett. Darpa timit acoustic-phonetic continous speech corpus cd-rom. nist speech disc 1-1.1. *NASA STI/Recon technical report n*, 93:27403, 1993.

Khaled Hechmi, Trung Ngo Trong, Ville Hautamäki, and Tomi Kinnunen. Voxceleb enrichment for age and gender recognition. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 687–693. IEEE, 2021.

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

Wei-Ning Hsu, Anuroop Sriram, Alexei Baevski, Tatiana Likhomanenko, Qiantong Xu, Vineel Pratap, Jacob Kahn, Ann Lee, Ronan Collobert, Gabriel Synnaeve, et al. Robust wav2vec 2.0: Analyzing domain shift in self-supervised pre-training. In *Proc. Interspeech 2021*, pages 721–725, 2021b.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

Shujie Hu, Long Zhou, Shujie Liu, Sanyuan Chen, Lingwei Meng, Hongkun Hao, Jing Pan, Xunying Liu, Jinyu Li, Sunit Sivasankaran, et al. Wavllm: Towards robust and adaptive speech large language model. *arXiv preprint arXiv:2404.00656*, 2024.

Mohammad Ali Humayun, Junaid Shuja, and Pg Emeroylariffion Abas. A review of social background profiling of speakers from speech accents. *PeerJ Computer Science*, 10:e1984, 2024.

Peter Kuppens, Francis Tuerlinckx, James A Russell, and Lisa Feldman Barrett. The relation between valence and arousal in subjective experience. *Psychological bulletin*, 139(4):917, 2013.

I Lawrence and Kuei Lin. A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, pages 255–268, 1989.

Colin Lea, Vikramjit Mitra, Aparna Joshi, Sachin Kajarekar, and Jeffrey P Bigham. Sep-28k: A dataset for stuttering event detection from podcasts with people who stutter. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6798–6802. IEEE, 2021.

Thanathai Lertpetchpun, Tiantian Feng, Dani Byrd, and Shrikanth Narayanan. Developing a high-performance framework for speech emotion recognition in naturalistic conditions challenge for emotional attribute prediction. In *Interspeech2025*, 2025.

Jingyi Li, Weiping Tu, and Li Xiao. Freevc: Towards high-quality text-free one-shot voice conversion. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

Reza Lotfian and Carlos Busso. Building naturalistic emotionally balanced speech corpus by retrieving emotional speech from existing podcast recordings. *IEEE Transactions on Affective Computing*, 10(4):471–483, 2017.

Dan Lyth and Simon King. Natural language guidance of high-fidelity text-to-speech with synthetic annotations. *arXiv preprint arXiv:2402.01912*, 2024.

Ziyang Ma, Zhisheng Zheng, Jiaxin Ye, Jinchao Li, Zhifu Gao, ShiLiang Zhang, and Xie Chen. emotion2vec: Self-supervised pre-training for speech emotion representation. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 15747–15760, 2024.

Brian MacWhinney. The talkbank project. In *Creating and digitizing language corpora: Volume 1: Synchronic databases*, pages 163–180. Springer.

Daniel Michelsanti, Zheng-Hua Tan, Shi-Xiong Zhang, Yong Xu, Meng Yu, Dong Yu, and Jesper Jensen. An overview of deep-learning-based audio-visual speech enhancement and separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 1368–1396, 2021.

Fatemehsadat Mireshghallah, Mohammadkazem Taram, Ali Jalali, Ahmed Taha Taha Elthakeb, Dean Tullsen, and Hadi Esmaeilzadeh. Not all features are equal: Discovering essential features for preserving prediction privacy. In *Proceedings of the Web Conference 2021*, pages 669–680, 2021.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*, 2017.

Tu Anh Nguyen, Wei-Ning Hsu, Antony D'Avirro, Bowen Shi, Itai Gat, Maryam Fazel-Zarani, Tal Remez, Jade Copet, Gabriel Synnaeve, Michael Hassid, et al. Expresso: A benchmark and analysis of discrete expressive speech resynthesis. In *Proc. Interspeech 2023*, pages 4823–4827, 2023.

OpenAI. Gpt-4 technical report. `https://openai.com/research/gpt-4`, 2024. Accessed: 2025-05-14.

Tae Jin Park, Naoyuki Kanda, Dimitrios Dimitriadis, Kyu J Han, Shinji Watanabe, and Shrikanth Narayanan. A review of speaker diarization: Recent advances with deep learning. *Computer Speech & Language*, 72:101317, 2022.

Leonardo Pepino, Pablo Riera, and Luciana Ferrer. Emotion recognition from speech using wav2vec 2.0 embeddings. In *Proc. Interspeech 2021*, pages 3400–3404, 2021.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, 2019.

Rohit Prabhavalkar, Takaaki Hori, Tara N Sainath, Ralf Schlüter, and Shinji Watanabe. End-to-end speech recognition: A survey. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:325–351, 2023.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023.

Nan Bernstein Ratner and Brian MacWhinney. Fluency bank: A new resource for fluency research and practice. *Journal of fluency disorders*, 56:69–80, 2018.

Dima Rekesh, Nithin Rao Koluguri, Samuel Kriman, Somshubra Majumdar, Vahid Noroozi, He Huang, Oleksii Hrinchuk, Krishna Puvvada, Ankur Kumar, Jagadeesh Balam, et al. Fast conformer with linearly scalable attention for efficient speech recognition. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE, 2023.

Ulrich Reubold, Jonathan Harrington, and Felicitas Kleber. Vocal aging effects on f0 and the first formant: A longitudinal analysis in adult speakers. *Speech communication*, 52 (7-8):638–651, 2010.

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. In *Proc. Interspeech 2024*, pages 4873–4877, 2024.

Donald L Rubin. Nonlanguage factors affecting undergraduates' judgments of nonnative english-speaking teaching assistants. *Research in Higher Education*, 33:511–531, 1992.

Ramon Sanabria, Nikolay Bogoychev, Nina Markl, Andrea Carmantini, Ondrej Klejch, and Peter Bell. The edinburgh international accents of english corpus: Towards the democratization of english asr. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

Susanne Schötz. *Perception, analysis and synthesis of speaker age*, volume 47. Lund University, 2006.

Shijing Si, Jianzong Wang, Junqing Peng, and Jing Xiao. Towards speaker age estimation with label distribution learning. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4618–4622. IEEE, 2022.

Sara Skoog Waller, Mårten Eriksson, and Patrik Sörqvist. Can you hear my age? influences of speech rate and speech spontaneity on estimation of speaker age. *Frontiers in psychology*, 6:978, 2015.

David Snyder, Daniel Garcia-Romero, Gregory Sell, Daniel Povey, and Sanjeev Khudanpur. X-vectors: Robust dnn embeddings for speaker recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5329–5333. IEEE, 2018.

Natalia Tomashenko, Xiaoxiao Miao, Pierre Champion, Sarina Meyer, Xin Wang, Emmanuel Vincent, Michele Panariello, Nicholas Evans, Junichi Yamagishi, and Massimiliano Todisco. The voiceprivacy 2024 challenge evaluation plan. *arXiv preprint arXiv:2404.02677*, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Christophe Veaux, Junichi Yamagishi, and Kirsten MacDonald. Cstr vctk corpus: English multi-speaker corpus for cstr voice cloning toolkit. *University of Edinburgh. The Centre for Speech Technology Research (CSTR)*, 2012.

Irina-Elena Veliche, Zhuangqun Huang, Vineeth Ayyat Kochaniyan, Fuchun Peng, Ozlem Kalinli, and Michael L Seltzer. Towards measuring fairness in speech recognition: Fair-speech dataset. In *Proc. Interspeech 2024*, pages 1385–1389, 2024.

Johannes Wagner, Andreas Triantafyllopoulos, Hagen Wierstorf, Maximilian Schmitt, Felix Burkhardt, Florian Eyben, and Björn W Schuller. Dawn of the transformer era in speech emotion recognition: closing the valence gap. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9):10745–10759, 2023.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003, 2021.

Wenbin Wang, Yang Song, and Sanjay Jha. Globe: A high-quality english corpus with global accents for zero-shot speaker adaptive text-to-speech. In *Proc. Interspeech 2024*, pages 1365–1369, 2024a.

Wenbin Wang, Yang Song, and Sanjay Jha. Usat: A universal speaker-adaptive text-to-speech approach. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024b.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, et al. Superb: Speech processing universal performance benchmark. In *Proc. Interspeech 2021*, pages 1194–1198, 2021.

Yuchen Yang, Thomas Thebaud, and Najim Dehak. Demographic attributes prediction from speech using wavlm embeddings. In *2025 59th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE, 2025.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246, 2018.

Ziqiang Zhang, Long Zhou, Chengyi Wang, Sanyuan Chen, Yu Wu, Shujie Liu, Zhuo Chen, Yanqing Liu, Huaming Wang, Jinyu Li, et al. Speak foreign languages with your own voice: Cross-lingual neural codec language modeling. *arXiv preprint arXiv:2303.03926*, 2023.

Guanlong Zhao, Sinem Sonsaat, Alif Silpachai, Ivana Lucic, Evgeny Chukharev-Hudilainen, John Levis, and Ricardo Gutierrez-Osuna. L2-arctic: A non-native english speech corpus. In *Proc. Interspeech 2018*, pages 2783–2787, 2018.

Jinzuomu Zhong, Korin Richmond, Zhiba Su, and Siqi Sun. Accentbox: Towards high-fidelity zero-shot accent generation. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

Juan Zuluaga-Gomez, Sara Ahmed, Danielius Visockas, and Cem Subakan. Commonaccent: Exploring large acoustic pretrained models for accent classification based on common voice. In *Proc. Interspeech 2023*, pages 5291–5295, 2023.