AFRINAMES: MOST ASR MODELS "BUTCHER" AFRICAN NAMES

Tobi Olatunji[1,13,*], **Tejumade Afonja**[2,3,10*], **Bonaventure F. P. Dossou** [4,5,6,7,*], **Atnafu Lambebo Tonja**[8,*], **Chris Chinenye Emezue**[5,7,9,*], **Amina Mardiyyah Rufai**[11,*], **Sahib Singh**[12,*]

[*]Masakhane NLP, [1] Intron Health Inc,[2]AI Saturdays Lagos,[3] CISPA Helmholtz Center for Information Security,
[4] Center for Intelligent Machines, McGill University, [5]Mila - Quebec AI Institute, [6] Lelapa AI, [7] Lanfrica,
[8] Instituto Politécnico Nacional, Mexico, [9] Technical University of Munich, [10] Saarland University,
[11] Idiap Research Institute, [12] Ford Motor Company, [13] Georgia Institute of Technology.

ABSTRACT

Useful conversational agents must accurately capture named entities to minimize error for downstream tasks, for example, asking a voice assistant to play a track from a certain artist, initiating navigation to a specific location, or documenting a diagnosis result for a specific patient. However, where named entities such as "Ukachukwu" (Igbo), "Lakicia" (Swahili), or "Ingabire" (Rwandan) are spoken, automatic speech recognition (ASR) models' performance degrades significantly, propagating errors to downstream systems. We model this problem as a distribution shift and demonstrate that such model bias can be mitigated through multilingual pre-training, intelligent data augmentation strategies to increase the representation of African-named entities, and fine-tuning multilingual ASR models on multiple African accents. The resulting fine-tuned models show an 86.4% relative improvement compared with the baseline on samples with African-named entities.

1 INTRODUCTION AND MOTIVATION

Automatic Speech Recognition (ASR) powers voice assistants, which use machine learning and other artificial intelligence techniques to automatically interpret and understand spoken languages for conversational purposes. With the advent of breakthroughs such as Google(Assistant, 2016), Amazon(Alexa, 2014), Apple(Siri, 2011), Samsung(Bixby, 2017), Microsoft(Cortana, 2014) etc., voice assistant technology has increasingly become a widespread technology (Siegert, 2021) with diverse applications which range from healthcare (Parente et al., 2004; Durling & Lumsden, 2008; Johnson et al., 2014), education(Bain et al., 2002; Wald, 2005) to other businesses applications(Li et al., 2015; Junqua & Haton, 2012). In recent years, the integration of conversational AI into smart devices has become common. However, with a growing user base from diverse demographics, there is a need for more inclusive and robust AI agents with better spoken language understanding (SLU) and accent recognition capabilities. (Press, 2022; Desot et al., 2019; Adelani et al., 2021)

Useful conversational agents must accurately capture named entities to minimize errors for downstream tasks. For example, in the command, "Play Billie Jean by Micheal Jackson", conversational agents need to excel at 3 core tasks: Speech Recognition, Named Entity Recognition, and Entity Linking, to appropriately respond to commands. The ASR component of the system must correctly transcribe the speech, laying a good foundation for Named Entity Recognition (NER) (Nguyen & Yu), which is, in turn, necessary for effective Entity Linking.

However, in the command "Play 'Trouble Sleep Yanga Wake Am' by Fela Anikulapo Kuti" (Fela is one of Africa's most legendary artists) spoken by a Nigerian with a thick Yoruba accent, the phonetic and linguistic variability of the heavily accented speech presents a double dilemma for such systems. Firstly, the heavy accent and tonality can be difficult for the system to recognize, and secondly, the use of out-of-vocabulary words can confuse the model, making it nearly impossible for the system to generate a correct response. Siri responds "I couldn't find 'trouble sleep younger we' by Fela and Kolapo Coochie in your library", typifying the failures of similar agents on out-of-distribution named entities. More examples are in Table 1.

Table 1: Model behavior examples on native African named entities

Model	Sentence
reference	**Ifeadigo** has been living at **Kaduna** with his wife **Chiamaka Orajimetochukwu**
azure	if you're diego.
gcp	diego has been living at his wife
aws	if you did good has been living at kaduna with his wife, she america or raji mo to
hubert-large-ls960-ft	ifia di gun has been living at cardonal with his wife shia maca orraji mo tu truku
hubert-xlarge-ls960-ft	ifia di gun ha been living at cardona with his wife shia macca or ragi mo tu truco
w2v2-lg-960h-lv60-self	fia digo has been living at cadna with his wife shi maca orajimo to truco o
w2v2-lg-960h	ifia digoun not been living at caduna with his wife shea macca or ra gi mou toul tru coul
w2v2-ft-swbd-300h	ifia digo has been living at cadona with his wife shi meca or ragimo too tro qo
w22-lg-xlsr-53-en	ifia digu has been living at kaduna with his wife shiamaka orajimo tutruku
w2v2-xls-r-1b-english	ifia digo has been living at kaduna with his wife shiamaka orajimu tutuku
whisper-base	if you are de-goon, that's when living at kaduna, which is wife, shia makka, or rajimu, to kuku
whisper-small	ifya digo has been living at kaduna with his wife, shiyamaka orajimu tochuku
whisper-medium	ifeia digun has been living at kaduna with his wife, shiamaka or rajimu, to chuku
whisper-large	ifeardigun has been living at kaduna with his wife, shiamaka or rajimu, to chukwu
xlsr-general (Ours)	ifiadigo has been living at kaduna with his wife chiamaka orajimotochukwu
Whisper-general (Ours)	ifeadigo has been living at kaduna with his wife chiamaka orajimotochukwu

We hypothesize that the under representation (and sometimes complete lack of) African named-entities in their training data may partly explain the model bias and eventual "butchering"[1] of African names by many voice assistants and conversational agents. We investigate state-of-the-art (SOTA) ASR models' performance on accented African speech with African and non-African named entities. Furthermore, we design a data augmentation strategy to increase the representation of African-named entities in speech corpora and demonstrate the effectiveness of our strategy through fine-tuning experiments on the augmented data.

Our contributions are as follows:

1. We investigate the performance of state-of-the-art ASR models on African named entities. To do this, we design an effective strategy to evaluate ASR models on speech datasets with no prior NER annotations. Our study highlights the failure of existing SOTA and commercial ASR models on samples with African named-entities
2. We develop a data augmentation strategy to increase the representation of African-named entities, creating a novel speech corpus rich in African named-entities, and show that by fine-tuning pre-trained models on the augmented accented data, we significantly improve the ability of pre-trained models to recognize African named entities. We open-source the dataset and fine-tuned models[2].

2 Related work

2.1 Improving monolingual and multilingual ASR for low-resource languages

The development of ASR systems for low-resource languages is still a challenging task due to the lack of training data and resources. Many researchers have reported a lack of generalization by models trained primarily on high-resource languages such as English, indicating a performance gap in ASR model's performance for low-resource languages (Lepak et al., 2021; Chuangsuwanich, 2016). As a

[1]To "butcher" a name means to mispronounce or incorrectly say someone's name. This term is often used to describe an error in pronunciation that makes the name sound significantly different from the correct pronunciation.

[2]https://huggingface.co/datasets/tobiolatunji/afrispeech-200

result, several approaches to overcome resource constraints are being developed in order to produce high-performing ASR models for low-resource languages. Such methods include cross-lingual representations, where the system learns a shared representation across multiple languages (Conneau et al., 2017), data augmentations techniques (Renduchintala et al., 2018; Feng et al., 2021; Fadaee et al., 2017), and leveraging high-resource language models fine-tuned on low-resource languages (Anaby-Tavor et al., 2020). The onset of multilingual speech corpora like Common Voice (Ardila et al., 2019b) led to the increased trend towards multilingual ASR and the release of large, robust multilingual pre-trained ASR models such as Whisper (96+ languages) (Radford et al., 2022) and Wav2vec-xls-r (128 languages) (Babu et al., 2022) along with several variations such as SpeechStew (Chan et al., 2021) and (Ritchie et al., 2022) trained on multilingual speech corpora. These models achieved SOTA performance on many downstream tasks, outperforming monolingual models like HuBERT (Hsu et al., 2021), wavLM (Chen et al., 2022), and wav2vec2 (Baevski et al., 2020), which were pre-trained or fine-tuned exclusively on monolingual corpora such as Librispeech (Panayotov et al., 2015), LibriVox (Kearns, 2014), SWBD (Godfrey et al., 1992), or WSJ (Paul & Baker, 1992). IndicWav2vec (Javed et al., 2021) showed that pretraining wav2vec on a large corpus of 40 Indian languages led to improved ASR performance on Indian ASR. Similar efforts by Faste (2022) for Irish ASR, Al-Ghezi et al. (2021) for L2 Swedish ASR, and Yi et al. (2020) achieved higher ASR performance in various low resource languages such as Mandarin, Japanese, Arabic, German.

2.2 Advances in accented English ASR

According to a survey of accented speech recognition (Hinsvark et al., 2021), the linguistic variability of accents presents hard challenges for ASR systems in both data collection and modeling strategies. Promising approaches include training accent-specific models (Vergyri et al., 2010; Najafian et al., 2014), data augmentation such as speed perturbation (Fukuda et al., 2018), model generalization through multi-task learning (Jain et al., 2018; Radford et al., 2022; Li et al., 2018), domain expansion (Ghorbani et al., 2019; Houston & Kirchhoff, 2020), pronunciation modification (Goronzy et al., 2004; Lehr et al., 2014), adaptation using auxiliary acoustic features (Grace et al., 2018; Li et al., 2018; Zhu et al., 2020), accent embeddings (Viglino et al., 2019; Turan et al., 2020), and adversarial training (Sun et al., 2018; Chen et al., 2020).

2.3 Racial Bias in ASR

Racial bias in ASR systems is an important problem that needs to be addressed to ensure that ASR systems are fair and accessible to all individuals, regardless of racial or ethnic background. A 2020 study by (Koenecke et al., 2020) found large racial disparities in the performance of five popular commercial ASR systems – Amazon, Apple, Google, IBM, and Microsoft – when transcribing structured interviews conducted with 42 white speakers and 73 black speakers. They found that all five ASR systems exhibited substantial racial disparities, with an average word error rate (WER) of 0.35 for black speakers compared with 0.19 for white speakers. This bias is representative of their underlying training data. Most ASR systems work best for native English speakers but their accuracy plummets drastically with non-native English speakers (Hassan et al., 2022; Prasad & Jyothi, 2020). Similarly, performance gaps with accented English have been well demonstrated by (Doumbouya et al., 2021; Siminyu et al., 2021; Babirye et al., 2022; Ogayo et al., 2022), with multiple parallel efforts (Gutkin et al., 2020; Dossou & Emezue, 2021; Afonja et al., 2021b; Kamper & Niesler, 2011) to create accented English datasets.

2.4 Improving ASR for named entities and Speech NER

Many studies have demonstrated progress and challenges in the field of NER for ASR and explored various methods to improve NER performance. French researchers (Galliano et al., 2009) outlined steps for evaluating NER on radio broadcast transcripts. (Xiao & Qian, 2021) in 2021 evaluated Chinese accented ASR on an Automatic Voice Query Service (AVQS), highlighting the severe limitation of such systems for multi-accented Mandarin users. They improve the final quality of the AVQS system by pairing an end-to-end Transformer-CTC ASR model with fuzzy logic (Xiao & Qian, 2021). (Katerenchuk & Rosenberg, 2014) and (Rangarajan & Narayanan, 2006) leverage prosodic features, (Ramabhadran et al., 2004) integrate document metadata to improve NER. More recently, (Mdhaffar et al., 2022; Caubrière et al., 2020) attempted to extract semantic information directly from speech signals with a single end-to-end model jointly learning ASR and NER tasks.

3 METHODOLOGY

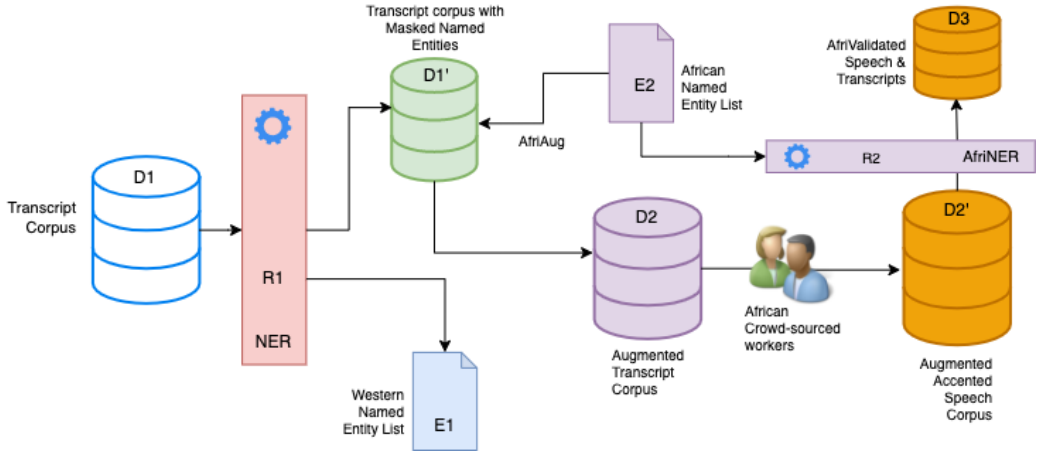

Figure 1: AfriNames dataset augmentation process.

Approach: We model the generalization problem as a domain shift and illustrate the workflow of the proposed solution in Figure 1. D_1 is a predominantly western dataset $\{(X^{E_1}, Y^{E_1})\}$ with audio and transcript pairs, originating from a distribution D^{E_1} induced by Anglo-centric named-entities E_1. Model M_1 is randomly initialized and trained on D_1, learning the mapping $f : X^{E_1} \longrightarrow Y^{E_1}$, leading to a pretrained model $M_1^{E_1}$. Pre-trained NER model R_1 extracts named entities from D_1 transcripts resulting in predominantly western named entities list E_1. Masking E_1 tokens from randomly selected samples in D_1 produces D_1'. Tokens from curated African named-entity list E_2 randomly replace masked tokens in D_1'. Augmented subset D_1' + D_1 creates D_2 transcripts which are sent to African crowd-sourced workers for recording to create D_2', a novel corpus of accented audio and augmented transcript pairs originating from a distribution D^{E_1} and D^{E_2} induced by African and Anglo-centric named-entities $E_1 + E_2 = E_3$. A Specialized NER model R_2 extracts African and western named entities E_3 from D_2 including any African named entities originally in D_1. Accented audio recordings of D_1' prompts and D_1' transcripts are isolated to create D_3, the subset of D_2' confirmed to contain African named entities (AfriValidated). A real-world example of D^{E_1} is LibriSpeech Panayotov et al. (2015), a 1,000-hours speech-text dataset from English-only audiobooks. The resulting ASR model $M_1^{E_1}$, such as Wav2vec2 Baevski et al. (2020), therefore, generalizes poorly to African named entities (Table 1).

3.1 DATASETS

In this study, we primarily explore the AfriSpeech-200 dataset, a 200hr novel accented English speech corpus rich with African-named entities, curated for clinical and general domain ASR. 67,577 prompts were recorded by 2,463 unique crowdsourced African speakers from 13 Anglophone countries across sub-Saharan Africa and the United States. The average audio duration was 10.7 seconds (see Table 2).

We explore two additional datasets (Table 3): (1) **SautiDB** (Afonja et al., 2021a), a dataset of Nigerian accent recordings with 919 audio samples, a sampling rate of 48kHz each, for a total amount of 59 minutes of recordings; (2) **Common Voice English Accented Dataset**, a subset of English Common Voice (version 10) (Ardila et al., 2019b) with majority American and European English accents removed.

3.2 AFROAUG: AFRICAN NAMED-ENTITY AUGMENTATION

Neural networks learn concepts from training data. Where transcripts or prompts in training data are predominantly Western (e.g. Common Voice and LibriSpeech (Ardila et al., 2019a; Panayotov et al., 2015)) and African-named entities are sparse, such ASR systems fail at correctly transcribing African

Table 2: AfriSpeech-200 Dataset statistics

	Train	Dev	Test
Duration (hrs)	173.4	8.74	18.77
# General domain clips	21,682	1,407	2,723
Unique Speakers	1,466	247	750
Accents	71	45	108
Average Audio duration	10.7 seconds		
Named Entities Category Counts			
PER	11,011	669	1,064
ORG	6,322	372	279
LOC	3,194	192	526

Speaker Gender Ratios - # Clip %	
Female	57.11%
Male	42.41%
Other/Unknown	0.48%
Speaker Age Groups - # Clips	
<18yrs	1,264 (1.88%)
19-25	36,728 (54.58%)
26-40	18,366 (27.29%)
41-55	10,374 (15.42%)
>56yrs	563 (0.84%)
Clip Domain - # Clips	
Clinical	41,765 (61.80%)
General	25,812 (38.20%)

names like "Ogochukwu" (Igbo), "Malaika" (Swahili), or "Uwimana" (Rwandan), while excellently transcribing Western names like "Lauren" and "Bryan"– representative of the bias in their training corpora. To increase the representation of African named entities, we start with a corpus D_1 using open-source predominantly western corpora: Wikitext-103 (Merity et al., 2016), Pubmed (Wheeler et al., 2007), and NCBI Disease (Doğan et al., 2014), and scrape African entertainment and news websites, and augment these datasets using two main strategies.

We derive a list E_2 of approximately 100k African names using a database of 90,000 African names from Anderson et al. (2013), 965 Nigerian Igbo names from Okagbue et al. (2017), and 1,000 African names obtained from freely available textbooks, online baby name websites, oral interviews, published articles, and online forums like Instagram and Twitter; and a list of African cities from Wikipedia. Next, we perform two key processes:

1. **Named-Entity Extraction with NER Models:** We leverage off-the-shelf pre-trained NER models (Conneau et al., 2019) and annotate all named-entities in corpus Y^{E_1} to extract the list E_1, tokens tagged with [PER], [LOC], or [ORG].

2. **Templating strategy:** Using randomly sampled transcripts y_i from Y^{E_1} with named entity tokens $e_i \in E_1$, we curate 140 template sentences and mask 1 to 4 named entities. We manually review, select and validate 140 transcripts where the replacement of masked tokens with African named-entities sound natural and retain their meaning in context. We randomly (uniformly) replace [LOC] tags with African cities from E_2, and [PER] and [ORG] tags with African names e_1 to e_4 from E_2. We repeat this process 200 times to create text corpus Y^{E_2} consisting of 28,000 augmented sentences from templates extracted from Y^{E_1} combined with the superset of sentences from Y^{E_1} (100,000+ sentences). We then give Y^{E_2} to crowd-sourced workers who record them, producing new audio samples X^{E_2}. In so doing, we create a new dataset $D_2 = \{(X^{E_2}, Y^{E_2})\}$ coming from a new distribution D^{E_2} induced by African entities E_2.

Given the new augmented training dataset D_2, the pre-trained ASR model $M_1^{E_1}$ learns a new mapping $f : X^{E_2} \longrightarrow Y^{E_2}$ during fine-tuning leading to a more robust fine-tuned model $M_1^{E_2}$, that is now able to adapt to the target distribution D^{E_2}.

Test Samples	#n	#entities	Categories		
			PER	ORG	LOC
AfriSpeech	2723	2478	1347	398	733
SautiDB	138	92	71	3	18
CV-En-Accented	334	170	76	38	56

Table 3: Dataset entity category counts.

3.3 AFRINER: NAMED-ENTITY EVALUATION

Given that D_2 is created by augmenting a subset of D_1, D_2, therefore, contains sentences with named entities from both E_1 and E_2. To evaluate NER on ASR-predicted transcripts from D_2, we need a reliable way to identify named entities in D_2 since D_2 has no prior ground truth NER annotations. To achieve this, we run NER inference on all test samples in D_2 using a specialized performant NER model[3] from (Adelani et al., 2022) that jointly predicts the set of African and western named entities E_3 in D_2.

3.4 AFRIVALIDATION: AFRICAN NAMED-ENTITY VALIDATION

To evaluate Named Entity Recognition (NER) specifically on African-named entities, we need to extract a subset of data from the larger dataset D_2 which contains both Western and African-named entities. This subset is called D_3 and it consists of samples from D_2 where African-named entities from E_2 exist. However, this process may not be perfect and some African names in D_2 may be missed, it ensures that we can evaluate a subset of test samples that contain African-named entities.

4 EXPERIMENTS

4.1 BENCHMARKS

We compare SOTA open-source pre-trained ASR models: Whisper (Radford et al., 2022), Wav2vec2 (Baevski et al., 2020), XLSR (Grosman, 2021), Hubert (Hsu et al., 2021), and WavLM (Chen et al., 2022), with commercial ASR systems. We refer readers to the respective papers for details on pre-training corpora, model architecture, and hyperparameters. We compare 3 model categories:

1. Models pre-trained or fine-tuned exclusively on predominantly Western transcripts, western English speech, and western named-entities
2. Models fine-tuned on predominantly Western transcripts, accented English speech, and predominantly western named entities
3. Commercial ASR APIs

4.2 FINE-TUNING

We select two best-performing open-source models from section 4.1, based on results in Table 4, and fine-tune them on an accented speech corpus dense with African and western-named entities to achieve robustness to western and African-named entities. We compare pre-trained model performance with fine-tuned checkpoints. Selected model architectures include:

- wav2vec2-large-xlsr-53 (Grosman, 2021): this model was fine-tuned on English using the train and validation split from (Ardila et al., 2019b), as detailed in (Conneau et al., 2020). It follows an encoder-decoder architecture with a CNN-based feature extractor, code book, and transformer-based encoder, 378.9M parameters; learning rate of 1e-4.
- whisper-medium (Radford et al., 2022): a decoder-only multi-task architecture, 789.9M parameters; learning rate of 2.5e-4. (We do not fine-tune whisper-large because of computational resource constraints)

For each model, we fine-tuned with FP16 Micikevicius et al. (2017), AdamW (Loshchilov & Hutter, 2017), batch size of 16, for 10 epochs, with a linear learning rate decay to zero after a warmup over the first 10% of iterations. XLSR was trained on a single Tesla T4 GPU with 16GB GPU memory while Whisper was trained on RTX8000 GPU with 48GB GPU memory. Fine-tuning took 2 days.

4.3 EVALUATION

Since ground truth NER labels do not exist for the selected speech datasets, we infer NER labels using a specialized NER model (Adelani et al., 2022) as described in section 3.4 above. We select

[3] https://huggingface.co/masakhane/afroxlmr-large-ner-masakhaner-1.0_2.0

Table 4: WER results for 4 model categories on test samples from 3 datasets - SautiDB (SDB) (Afonja et al., 2021a), CommonVoice-En (CV) (Ardila et al., 2019a) using an accented subset and AfriSpeech (Afri). The table shows selected models, the number of parameters, number of fine-tuning corpora ("Multi" refers to multilingual or multi-task pre-training corpora); $ents$ represent the number of named entity instances [PER + ORG + LOC] in the test data, n represents the total number of test samples, p is the number of samples in n with no named entities m is the number of samples in n with named-entities, r is the number of samples in m with African-named entities. It follows that $r < m < p < n$. **All** represents mean WER across all test samples in the dataset. **No-NER** represents mean WER across sentences with NO predicted named entities. **NER** represents mean WER across all sentences WITH predicted named entities. **AfriV** represents mean WER across the AfriValidated sentences where African-named entities from E_2 exist.

Model	Params	#ents	WER							
			#n	All	#p	No-NER	#m	NER	#r	AfriV
Baseline										
wav2vec2-large-960h	317M	2478	2723	0.641	1257	0.575	1215	0.696	117	0.721
Monolingual Fine-tuning: Open-Source SOTA pre-trained Models										
wav2vec2-lg-960h-lv60-self	317M	2478	2723	0.533	1257	0.469	1215	0.584	117	0.601
hubert-large-ls960-ft	317M	2478	2723	0.557	1257	0.494	1215	0.607	117	0.613
hubert-xlarge-ls960-ft	317M	2478	2723	0.562	1257	0.498	1215	0.613	117	0.627
wavlm-libri-clean-100h-large	317M	2478	2723	0.631	1257	0.573	1215	0.680	117	0.685
w2v2-lg-robust-swbd-300h	317M	2478	2723	0.733	1257	0.660	1215	0.796	117	0.844
Multilingual Fine-tuning: Open-Source SOTA pre-trained Models										
whisper-large	1550M	2478	2723	0.240	1257	0.185	1215	0.300	117	0.383
whisper-medium	769M	2478	2723	0.276	1257	0.205	1215	0.352	117	0.420
whisper-small	244M	2478	2723	0.330	1257	0.257	1215	0.405	117	0.462
wav2vec2-lg-xlsr-53-english	317M	2478	2723	0.506	1257	0.457	1215	0.550	117	0.548
wav2vec2-xls-r-1b-english	317M	2478	2723	0.521	1257	0.468	1215	0.568	117	0.581
Commercial ASR APIs										
Azure	-	2478	2723	0.340	1257	0.280	1215	0.402	117	0.483
GCP	-	2478	2723	0.534	1257	0.467	1215	0.603	117	0.676
AWS	-	2478	2723	0.354	1257	0.281	1215	0.426	117	0.502
Western named entities fine-tuned on accented speech										
xlsr-53-english-SautiDB	317M	92	138	0.5	48	-	90	0.157	0	-
xlsr-53-english-CV	317M	170	334	0.3	190	-	144	0.253	3	-
AfriSpeech Finetuning Ours										
whisper-medium-AfriSpeech	769M, AfriSpeech	2478	2723	**0.186**	1257	**0.169**	1215	**0.198**	117	**0.098**
xlsr-53-english-AfriSpeech	769M, AfriSpeech	2478	2723	0.236	1257	0.209	1215	0.258	117	0.184

test sentences where an entity is detected with confidence (score) greater than 0.8 (this seemed to be a reasonable threshold based on ad-hoc analysis) and evaluate using Word Error Rate (WER). For selected pre-trained, commercial ASR models, and fine-tuned models, we evaluate WER on samples containing one or more named entities and present single-run results in Table 4.

5 Results and Discussion

5.1 African named entities are challenging

The baseline model[4] in Table 4 demonstrates the dominant trend in our results. WER on all samples (column 7, All) improves by 10.3% (relative) when samples with named entities are EXCLUDED (column 9, No-NER), worsens by 8.6% (relative) when samples with named entities (western + African) are isolated (column 11, NER). Performance sinks by 12.5% (relative) on the subset of Afrivalidated examples (column 13, AfriV)– samples with African-named entities. This pattern is consistent across all model categories except Ours where there is a 47.3% (whisper) and 22% (xlsr) relative WER improvement on Afrivalidated sentences.

5.2 Training data bias

As shown in Table 1 and 4, multilingual/multitask pre-training outperforms monolingual pre-training/fine-tuning. Multilingual/multitask models (Radford et al., 2022; Grosman, 2021; Gulati et al., 2020) learn more useful representations, are more linguistically diverse, robust, and generalize

[4]We considered wav2vec2-large-960h as a baseline since it is primarily trained on western data, e.g librispeech.

better to accented speech when compared to models fine-tuned extensively on Librispeech (Panayotov et al., 2015) and Switchboard (Godfrey et al., 1992), monolingual speech corpora with predominantly western accents. After fine-tuning the AfriSpeech data with African-named entities, our best model, whisper-medium-general improves on the baseline by 86.4%. Mean WER per model category: Monolingual 0.681; Multilingual 0.479; Commercial 0.554; Ours 0.141 (80.4% improvement over baseline).

5.3 NER MODELS AND AFRIVALIDATION

Figure 2 shows the distribution of named entities in the AfriSpeech dataset. Although NER models are imperfect, we manually validated several outputs that gave us confidence that the specialized NER model (Adelani et al., 2022) was in fact correctly identifying African and Western-named entities. Afrivalidation also guaranteed sentences with African-named entities were isolated. However, there are caveats. The African slave name database Anderson et al. (2013) contained Western names like George and John which were initially picked up during Afrivalidation. To mitigate this, we limit Afrivalidation to a set of nearly 2k Nigerian names described in section 3.2. Additionally, we found inconsistencies in the medical domain NER prediction, including wrongly recognized medical terms as named entities and incorrect prediction of non-[ORG] tokens as [ORG] with high confidence. To ensure a fair comparison, we focus on the general domain corpus and only investigate sentences with at least [PER] or [LOC] tags present. Several [PER], [LOC], and [ORG] examples can be seen in Appendix Tables 6 and 5.

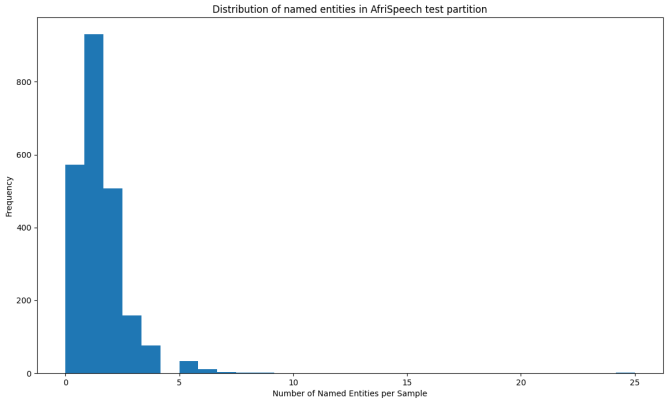

Figure 2: Distribution of named entities in AfriSpeech test partition

5.4 MULTILINGUAL PRE-TRAINING IS NOT SUFFICIENT

Appendix Tables 6 and 5 show several examples with African-named entities, comparing pre-trained vs fine-tuned versions of our best model, Whisper-medium. The finetuned model outperforms the pre-trained model by 76.67% relative. These results demonstrate that performant multilingual/multi-task models pre-trained on 90+ languages make several mistakes with African-named entities. Fine-tuning results show that our approach is effective in mitigating bias in these large models. Appendix Table 6 shows Afrivalidated named entity examples where the pre-trained whisper had a WER < 0.20 (low error) along with the improved transcripts after fine-tuning. Appendix Table 5 shows high error (WER > 0.8) examples along with improved transcripts by the fine-tuned whisper model.

5.5 USE OF LANGUAGE MODELS

Table 1 shows some of the difficulties with commercial APIs where a language model (LM) is likely used to refine or autocorrect the raw ASR predictions. This is especially damaging for African-named entities. Because these named entities (e.g."Ifeadigo") are missing from LM training data, the probability of sequences with these words is effectively zero, and such transcripts are downranked by

the algorithm in favor of more likely tokens like "Diego" as seen in the example in Table 1. Prediction score thresholds may also be in use under the hood in these commercial systems, limiting the ASR output where confidence is low resulting in truncated output as seen in Table 1.

6 CONCLUSION

Automatic speech recognition (ASR) for African-named entities is a challenging task for most state-of-the-art (SOTA) ASR models including those trained with multilingual data and multitask objectives. We demonstrate that this bias can be mitigated by fine-tuning these models on accented speech corpora rich in African-named entities, a distribution shift that improves their robustness in the African context.

7 LIMITATION AND FUTURE WORK

The appendix provides examples where the pre-trained model can recognize some named entities even without explicit training, as well as cases where African named entities are not recognized by the fine-tuned model. It is important to note that while adding an African named entity dataset is a positive step, incorporating it into the training process is necessary for the model to explicitly learn these entities.

ACKNOWLEDGMENTS

Tobi Olatunji acknowledges Intron Health Inc for providing the dataset and compute resources. Chris Chinenye Emezue acknowledges the support of the Mila - Quebec AI Institute for compute resources. Tejumade Afonja acknowledges the support of ELSA - European Lighthouse on Secure and Safe AI[5] in the form of an international travel and accommodation grant, which enabled her to attend the conference.

REFERENCES

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, et al. Masakhaner: Named entity recognition for african languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131, 2021.

David Ifeoluwa Adelani, Graham Neubig, Sebastian Ruder, Shruti Rijhwani, Michael Beukman, Chester Palen-Michel, Constantine Lignos, Jesujoba Oluwadara Alabi, Shamsuddeen Hassan Muhammad, Peter Nabende, Cheikh M. Bamba Dione, Andiswa Bukula, Rooweither Mabuya, Bonaventure F. P. Dossou, Blessing K. Sibanda, Happy Buzaaba, Jonathan Mukiibi, Godson Kalipe, Derguene Mbaye, Amelia Taylor, Fatoumata Kabore, Chris C. Emezue, Anuoluwapo Aremu, Perez Ogayo, Catherine W. Gitau, Edwin Munkoh-Buabeng, Victoire Memdjokam Koagne, Allahsera Auguste Tapo, Tebogo Macucwa, Vukosi Marivate, Elvis Mboning, Tajuddeen R. Gwadabe, Tosin P. Adewumi, Orevaoghene Ahia, Joyce Nakatumba-Nabende, Neo L. Mokono, Ignatius M Ezeani, Chiamaka Ijeoma Chukwuneke, Mofetoluwa Adeyemi, Gilles Hacheme, Idris Abdulmumin, Odunayo Ogundepo, Oreen Yousuf, Tatiana Moteu Ngoli, and Dietrich Klakow. Masakhaner 2.0: Africa-centric transfer learning for named entity recognition. *ArXiv*, abs/2210.12391, 2022.

Tejumade Afonja, Clinton Mbataku, Ademola Malomo, Olumide Okubadejo, Lawrence Francis, Munachiso Nwadike, and Iroro Orife. Sautidb: Nigerian accent dataset collection, 2021a.

Tejumade Afonja, Oladimeji Mudele, Iroro Orife, Kenechi Dukor, Lawrence Francis, Duru Goodness, Oluwafemi Azeez, Ademola Malomo, and Clinton Mbataku. Learning nigerian accent embeddings from speech: preliminary results based on sautidb-naija corpus. *arXiv preprint arXiv:2112.06199*, 2021b.

[5]ELSA is funded by the European Union under grant agreement No. 101070617. However, the views expressed in this work are solely those of the authors and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for any opinions or conclusions drawn from this research.

Ragheb Al-Ghezi, Yaroslav Getman, Aku Rouhe, Raili Hildén, and Mikko Kurimo. Self-supervised end-to-end asr for low resource l2 swedish. In *22nd Annual Conference of the International Speech Communication Association, INTERSPEECH 2021*. ISCA, 2021.

Amazon Alexa. `https://alexa.amazon.com/`, 2014.

Ateret Anaby-Tavor, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling. Do not have enough data? deep learning to the rescue! In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 7383–7390, 2020.

Richard Anderson, Alex Borucki, Daniel Domingues Da Silva, David Eltis, Paul Lachance, Philip Misevich, and Olatunji Ojo. Using african names to identify the origins of captives in the transatlantic slave trade: crowd-sourcing and the registers of liberated africans, 1808–1862. *History in Africa*, 40(1):165–191, 2013.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*, 2019a.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*, 2019b.

Google Assistant. `https://assistant.google.com/`, 2016.

Claire Babirye, Joyce Nakatumba-Nabende, Andrew Katumba, Ronald Ogwang, Jeremy Tusubira Francis, Jonathan Mukiibi, Medadi Ssentanda, Lilian D Wanzare, and Davis David. Building text and speech datasets for low resourced languages: A case of languages in east africa. In *3rd Workshop on African Natural Language Processing*, 2022.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Miguel Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. Xls-r: Self-supervised cross-lingual speech representation learning at scale. In *INTERSPEECH*, 2022.

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.

Keith Bain, Sara H Basson, and Mike Wald. Speech recognition in university classrooms: Liberated learning project. In *Proceedings of the fifth international ACM conference on Assistive technologies*, pp. 192–196, 2002.

Samsung Bixby. `https://www.samsung.com/us/apps/bixby/`, 2017.

Antoine Caubrière, Sophie Rosset, Yannick Estève, Antoine Laurent, and Emmanuel Morin. Where are we in named entity recognition from speech? In *Proceedings of the 12th Language Resources and Evaluation Conference*, pp. 4514–4520, 2020.

William Chan, Daniel Park, Chris Lee, Yu Zhang, Quoc Le, and Mohammad Norouzi. Speechstew: Simply mix all available speech recognition data to train one large neural network. *arXiv preprint arXiv:2104.02133*, 2021.

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.

Yi-Chen Chen, Zhaojun Yang, Ching-Feng Yeh, Mahaveer Jain, and Michael L Seltzer. Aipnet: Generative adversarial pre-training of accent-invariant networks for end-to-end speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6979–6983. IEEE, 2020.

Ekapol Chuangsuwanich. Multilingual techniques for low resource automatic speech recognition. Technical report, Massachusetts Institute of Technology Cambridge United States, 2016.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*, 2017.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*, 2020.

Microsoft Cortana. `https://www.microsoft.com/en-us/cortana`, 2014.

Thierry Desot, François Portet, and Michel Vacher. Towards end-to-end spoken intent recognition in smart home. In *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pp. 1–8. IEEE, 2019.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10, 2014.

Bonaventure FP Dossou and Chris C Emezue. Okwugb\'e: End-to-end speech recognition for fon and igbo. *arXiv preprint arXiv:2103.07762*, 2021.

Moussa Doumbouya, Lisa Einstein, and Chris Piech. Using radio archives for low-resource speech recognition: towards an intelligent virtual assistant for illiterate users. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 14757–14765, 2021.

Scott Durling and Jo Lumsden. Speech recognition use in healthcare applications. In *Proceedings of the 6th international conference on advances in mobile computing and multimedia*, pp. 473–478, 2008.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. Data augmentation for low-resource neural machine translation. *arXiv preprint arXiv:1705.00440*, 2017.

Sarah Faste. *WAV2VEC 2.0 FOR IRISH ASR: A MULTILINGUAL APPROACH TO UNDER-RESOURCED LANGUAGES*. PhD thesis, 2022.

Steven Y Feng, Varun Gangal, Jason Wei, Sarath Chandar, Soroush Vosoughi, Teruko Mitamura, and Eduard Hovy. A survey of data augmentation approaches for nlp. *arXiv preprint arXiv:2105.03075*, 2021.

Takashi Fukuda, Raul Fernandez, Andrew Rosenberg, Samuel Thomas, Bhuvana Ramabhadran, Alexander Sorin, and Gakuto Kurata. Data augmentation improves recognition of foreign accented speech. In *Interspeech*, number September, pp. 2409–2413, 2018.

Sylvain Galliano, Guillaume Gravier, and Laura Chaubard. The ester 2 evaluation campaign for the rich transcription of french radio broadcasts. In *Tenth Annual Conference of the International Speech Communication Association*, 2009.

Shahram Ghorbani, Soheil Khorram, and John HL Hansen. Domain expansion in dnn-based acoustic models for robust speech recognition. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 107–113. IEEE, 2019.

John J Godfrey, Edward C Holliman, and Jane McDaniel. Switchboard: Telephone speech corpus for research and development. In *Acoustics, Speech, and Signal Processing, IEEE International Conference on*, volume 1, pp. 517–520. IEEE Computer Society, 1992.

Silke Goronzy, Stefan Rapp, and Ralf Kompe. Generating non-native pronunciation variants for lexicon adaptation. *Speech Communication*, 42(1):109–123, 2004.

Mikaela Grace, Meysam Bastani, and Eugene Weinstein. Occam's adaptation: A comparison of interpolation of bases adaptation methods for multi-dialect acoustic modeling with lstms. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pp. 174–181. IEEE, 2018.

Jonatas Grosman. Fine-tuned XLSR-53 large model for speech recognition in English. `https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-english`, 2021.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*, 2020.

Alexander Gutkin, Isin Demirsahin, Oddur Kjartansson, Clara E Rivera, and Kólá Túbòsún. Developing an open-source corpus of yoruba speech. 2020.

Muhammad Ahmed Hassan, Asim Rehmat, Muhammad Usman Ghani Khan, Muhammad Haroon Yousaf, et al. Improvement in automatic speech recognition of south asian accent using transfer learning of deepspeech2. *Mathematical Problems in Engineering*, 2022, 2022.

Arthur Hinsvark, Natalie Delworth, Miguel Del Rio, Quinten McNamara, Joshua Dong, Ryan Westerman, Michelle Huang, Joseph Palakapilly, Jennifer Drexler, Ilya Pirkin, et al. Accented speech recognition: A survey. *arXiv preprint arXiv:2104.10747*, 2021.

Brady Houston and Katrin Kirchhoff. Continual learning for multi-dialect acoustic models. 2020.

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, 2021.

Abhinav Jain, Minali Upreti, and Preethi Jyothi. Improved accented speech recognition using accent embeddings and multi-task learning. In *Interspeech*, pp. 2454–2458, 2018.

Tahir Javed, Sumanth Doddapaneni, Abhigyan Raman, Kaushal Santosh Bhogale, Gowtham Ramesh, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. Towards building asr systems for the next billion users, 2021. URL `https://arxiv.org/abs/2111.03945`.

Maree Johnson, Samuel Lapkin, Vanessa Long, Paula Sanchez, Hanna Suominen, Jim Basilakis, and Linda Dawson. A systematic review of speech recognition technology in health care. *BMC medical informatics and decision making*, 14(1):1–14, 2014.

Jean-Claude Junqua and Jean-Paul Haton. *Robustness in automatic speech recognition: fundamentals and applications*, volume 341. Springer Science & Business Media, 2012.

Herman Kamper and Thomas Niesler. Multi-accent speech recognition of afrikaans, black and white varieties of south african english. In *Twelfth Annual Conference of the International Speech Communication Association*, 2011.

Denys Katerenchuk and Andrew Rosenberg. Improving named entity recognition with prosodic features. In *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

Jodi Kearns. Librivox: Free public domain audiobooks. *Reference Reviews*, 28(1):7–8, 2014.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R Rickford, Dan Jurafsky, and Sharad Goel. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689, 2020.

Maider Lehr, Kyle Gorman, and Izhak Shafran. Discriminative pronunciation modeling for dialectal speech recognition. 2014.

Łukasz Lepak, Kacper Radzikowski, Robert Nowak, and Karol J Piczak. Generalisation gap of keyword spotters in a cross-speaker low-resource scenario. *Sensors*, 21(24):8313, 2021.

Bo Li, Tara N Sainath, Khe Chai Sim, Michiel Bacchiani, Eugene Weinstein, Patrick Nguyen, Zhifeng Chen, Yanghui Wu, and Kanishka Rao. Multi-dialect speech recognition with a single sequence-to-sequence model. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 4749–4753. IEEE, 2018.

Jinyu Li, Li Deng, Reinhold Haeb-Umbach, and Yifan Gong. Robust automatic speech recognition: a bridge to practical applications. 2015.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Salima Mdhaffar, Jarod Duret, Titouan Parcollet, and Yannick Estève. End-to-end model for named entity recognition from speech without paired training data. *arXiv preprint arXiv:2204.00803*, 2022.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models, 2016.

Paulius Micikevicius, Sharan Narang, Jonah Alben, Gregory Diamos, Erich Elsen, David Garcia, Boris Ginsburg, Michael Houston, Oleksii Kuchaiev, Ganesh Venkatesh, et al. Mixed precision training. *arXiv preprint arXiv:1710.03740*, 2017.

Maryam Najafian, Andrea DeMarco, Stephen Cox, and Martin Russell. Unsupervised model selection for recognition of regional accented speech. In *Fifteenth annual conference of the international speech communication association*, 2014.

Minh Nguyen and Zhou Yu. Improving named entity recognition in spoken dialog systems by context and speech pattern modeling. URL https://aclanthology.org/2021.sigdial-1.6.

Perez Ogayo, Graham Neubig, and Alan W Black. Building african voices. *arXiv preprint arXiv:2207.00688*, 2022.

Hilary I Okagbue, Abiodun A Opanuga, Muminu O Adamu, Paulinus O Ugwoke, Emmanuela CM Obasi, and Grace A Eze. Personal name in igbo culture: a dataset on randomly selected personal names and their statistical analysis. *Data in brief*, 15:72–80, 2017.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5206–5210. IEEE, 2015.

Ronaldo Parente, Ned Kock, and John Sonsini. An analysis of the implementation and impact of speech-recognition technology in the healthcare sector. *Perspectives in Health Information Management/AHIMA, American Health Information Management Association*, 1, 2004.

Douglas B Paul and Janet Baker. The design for the wall street journal-based csr corpus. In *Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992*, 1992.

Archiki Prasad and Preethi Jyothi. How accents confound: Probing for accent information in end-to-end speech recognition systems. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 3739–3753, Online, July 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.acl-main.345.

Tsinghua University Press. Researchers propose new and more effective model for automatic speech recognition, Sep 2022. URL https://techxplore.com/news/2022-09-effective-automatic-speech-recognition.html.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*, 2022.

Bhuvana Ramabhadran, Olivier Siohan, and Geoffrey Zweig. Use of metadata to improve recognition of spontaneous speech and named entities. In *Eighth International Conference on Spoken Language Processing*, 2004.

Vivek Rangarajan and Shrikanth Narayanan. Detection of non-native named entities using prosodic features for improved speech recognition and translation. In *Multilingual Speech and Language Processing*, 2006.

Adithya Renduchintala, Shuoyang Ding, Matthew Wiesner, and Shinji Watanabe. Multi-modal data augmentation for end-to-end asr. *arXiv preprint arXiv:1803.10299*, 2018.

Sandy Ritchie, You-Chi Cheng, Mingqing Chen, Rajiv Mathews, Daan van Esch, Bo Li, and Khe Chai Sim. Large vocabulary speech recognition for languages of africa: multilingual modeling and self-supervised learning. *arXiv preprint arXiv:2208.03067*, 2022.

Ingo Siegert. Speaker anonymization solution for public voice-assistant interactions–presentation of a work in progress development. In *Proc. 2021 ISCA Symposium on Security and Privacy in Speech Communication*, pp. 80–82, 2021.

Kathleen Siminyu, Godson Kalipe, Davor Orlic, Jade Abbott, Vukosi Marivate, Sackey Freshia, Prateek Sibal, Bhanu Neupane, David I Adelani, Amelia Taylor, et al. Ai4d–african language program. *arXiv preprint arXiv:2104.02516*, 2021.

Apple Siri. `https://www.apple.com/siri/`, 2011.

Sining Sun, Ching-Feng Yeh, Mei-Yuh Hwang, Mari Ostendorf, and Lei Xie. Domain adversarial training for accented speech recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4854–4858. IEEE, 2018.

Mehmet Ali Tuğtekin Turan, Emmanuel Vincent, and Denis Jouvet. Achieving multi-accent asr via unsupervised acoustic model adaptation. In *INTERSPEECH 2020*, 2020.

Dimitra Vergyri, Lori Lamel, and Jean-Luc Gauvain. Automatic speech recognition of multiple accented english data. In *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

Thibault Viglino, Petr Motlicek, and Milos Cernak. End-to-end accented speech recognition. In *Interspeech*, pp. 2140–2144, 2019.

Mike Wald. Using automatic speech recognition to enhance education for all students: Turning a vision into reality. In *Proceedings Frontiers in Education 35th Annual Conference*, pp. S3G–S3G. IEEE, 2005.

David L Wheeler, Tanya Barrett, Dennis A Benson, Stephen H Bryant, Kathi Canese, Vyacheslav Chetvernin, Deanna M Church, Michael DiCuccio, Ron Edgar, Scott Federhen, et al. Database resources of the national center for biotechnology information. *Nucleic acids research*, 36(suppl_1): D13–D21, 2007.

Kejing Xiao and Zhaopeng Qian. Automatic voice query service for multi-accented mandarin speech. In *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pp. 2875–2881. IEEE, 2021.

Cheng Yi, Jianzhong Wang, Ning Cheng, Shiyu Zhou, and Bo Xu. Applying wav2vec2. 0 to speech recognition in various low-resource languages. *arXiv preprint arXiv:2012.12121*, 2020.

Han Zhu, Li Wang, Pengyuan Zhang, and Yonghong Yan. Multi-accent adaptation based on gate mechanism. *arXiv preprint arXiv:2011.02774*, 2020.

A Appendix

Table 5: Examples using our best (fine-tuned) model, Whisper-Medium-General, showing samples with WER > 0.8 on the pre-trained whisper model

PER	LOC	ORG	Reference	Prediction Pre-trained	WER Pre-trained	Prediction Fine-tune	WER Fine-tune
daberechi , iniola	-	-	dr daberechi neonatal intensive care unit (icu) aware and dr iniola surgery notified. 09 january, 2003	dr. davirechi nyunato, intensive care unit, awuya, and dr. inuyo la sajar, notified 9th january, 2003.	0.813	dr daberechi neonatal intensive care unit (icu) and dr inyola surgery notified. 09 jan, 2003	0.188
uloaku, onyinyechukwu, ne, ice finidi	-	-	dr. uloaku is w/ the pt onyinyechukwu at this time and has also spoken to pt's neice finidi	the other first of uno aco is w- the patient on iyukchuku at the time and has also spoken to patient apostrophe s, ms. findi.	0.889	dr. nnuaku is w/ the pt onyinyechukwu at the time and has also spoken to pt's neice tinde	0.167
udeme obong, abiona	ikeja	-	udeme obong will be in to visit abiona in the am at ikeja	in 2 days, ubuntu will be introduced in the am or tk job.	0.846	udimi obong will be in to visit obiona in the am for thickly joe	0.385
zeribe	-	-	patient zeribe presented on account of ammenorrhea of 4 months. next line. hot flushes associated with night sweats	patient 0 will be represented on a count of arm and ear of a 4-month-old. next line. outflotches are associated with 9th sweat.	0.833	patient zirinbe presented on account of ammenorrhea of 4 months. next line. hot flushes associated with night sweats	0.0556
ubanwa ibimina, kasiemobi	-	-	thu 04 feb, 1988 ob: dr. ubanwa ibimina dr. kasiemobi	thursday, 4th february, 1988, obi-kolun, dr. ubangwa ibimina, dr. kasiem obi,	0.900	thursday 04 february, 1988 ob: dr. obanwa ibimina dr. kasiemobi	0.300
ogechukwukana, mahaja onyedikachukwu	birnin kebbi	-	ogechukwukana has been living at birnin kebbi with his wife mahaja onyedikachukwu who helps with his medications.	so,	1.000	ogichukwukana has been living at birnin kebbi with his wife mahaja oyedikachukwu who helps with his medications.	0.118
chiehidra nkediniruka , obi	warri		top gynecologists, drs chiehidra nkediniruka and obi at warri leading specialist hospital were quizzed on the management of several patients.	i'm a college is common that does she i'd rather in getting a little bit for really in specialist hospital yes kids on the management of several patients who stop	1.150	top gynecologists, drs chiehidra nketeenerewa and obi at warri leading specialist hospital were skilled to the management of several patients.	0.150
ibiama', ayotola, nnennaya	potiskum	lekki clinic	patient's family members ibiama and ayotola showed up to the potiskum ward this morning looking for nnennaya who passed away last night at lekki clinic.	patient family members	0.920	patient's family members yabiamma and ayotola showed up to the potiskum ward this morning looking for nnennaya who passed away last night at 30 clinic.	0.080

Table 6: Examples using our best (fine-tuned) model, Whisper-Medium-General, showing samples with WER < 0.2 on the pre-trained whisper model

PER	LOC	ORG	Reference	Prediction Pre-trained	WER Pre-trained	Prediction Fine-tune	WER Fine-tune
femi	nigeria	-	femi says 21 not 18 persons have been killed in the first 14 days of the coronavirus lockdown in nigeria so far.	phenyl says 21 not 18 persons have been cured in the first 14 days of the coronavirus lockdown in nigeria so far.	0.091	femi says 21 not 18 persons have been killed in the first 14 days of the coronavirus lockdown in nigeria so far.	0.000
chinweizu Ojo, bu, kola	eket	-	children chinweizu ojo and bukola were found last night wandering the streets unattended after their mother and father, went missing while returning from work at eket	children chinweuzu, ojo and bukola were found last night wandering the streets unattended after their mother and father, went missing while returning from work at eket.	0.077	children chinweizu ojo and bukola were found last night wandering the streets unattended after their mother and father, went missing while returning from work at eket	0.000
N, nanna, Chimaihe	-	-	nnanna was watching tv as they normally do in the evening when his brother chimaihe went to prepare dinner.	nana was watching tv as they normally do in the evening when his brother chimaehe went to prepare dinner.	0.105	nnanna was watching tv as they normally do in the evening when his brother chimaihe went to prepare dinner.	0.000
femi	-	-	the event, which lasted five hours, was also used as a platform to surprise femi with a rare saxophone as a birthday gift before he sets out on his seven-week american tour.	the event, which lasted five hours, was also used as a platform to surprise femi with a real saxophone as a birthday gift before it set out on a seven-week american tour.	0.125	the event, which lasted five hours, was also used as a platform to surprise femi with a rare saxophone as a birthday gift before he sets out on his seven week american tour.	0.0625
akubuilo	-	umuahia elementary school	akubuilo began playing the piano when he was a young child at umuahia elementary school	akubuilo began playing the piano when he was a young child at omuahe elementary school.	0.133	acobuilo began playing the piano when he was a young child at omuahia elementary school	0.133
kilani	-	asaba elementary school	kilani began playing the piano when he was a young child at asaba elementary school	killani began playing the piano when he was a young child at asaba elementary school.	0.133	kilani began playing the piano when he was a young child at asaba elementary school	0.000
ihuoma , inango	-	-	patient ihuoma was addicted to morphine and eventually had to see dr. inango	patient ioma was addicted to morphine and eventually had to see dr. enango .	0.154	patient ihuoma was addicted to morphine and eventually had to see dr. inango	0.000
ojo	jordan	nigerian tribune	speaking with the nigerian tribune on the significance of jordan to the christian faithful, prophet ojo described it as a land of separation from the land of reproach to that of promise.	speaking with the nigerian tribune on the significance of jordan to the christian faithful, prophet ojo described it as a land of separation from the land of reproach to that of promise.	0.000	speaking with the nigerian tribune on the significance of jordan to the christian faithful, prophet ojoe described it as a land of separation from the land of rep ridge to that of promise.	0.0938