INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

AfriNames: Most ASR models "butcher" African Names

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

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

tobi@intron.io, tejumade.afonja@cispa.de, bonaventure.dossou@mila.quebec, atnafu.lambebo@wsu.edu.et, chris.emezue@gmail.com, amina.rufai@idiap.ch, sahibsingh570@gmail.com

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 laboratory result for a 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 81.5% relative WER improvement compared with the baseline on samples with African-named entities.

Index Terms: Speech recognition, named entity recognition, distribution shift, accented speech

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 Amazon's Alexa, and Apple's Siri, etc., voice assistant technology has increasingly become a widespread technology with diverse applications [1]. However, as these devices gain adoption beyond the demographics of their training data, there is a need for more inclusive and robust AI agents with better spoken language understanding (SLU) and accent recognition capabilities [2, 3][1].

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) [4], which is, in turn, necessary for effective Entity Linking.

However, in the command "Play 'Trouble Sleep Yanga Wake Am' by Fela Anikulapo Kuti"[2] 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", effectively "butchering" [3] the name, typifying the failures of similar agents on out-of-distribution named entities. More examples in Table 1.

We hypothesize that the underrepresentation (and sometimes complete lack of) African named-entities in their training data may partly explain the model bias and eventual "butchering" of African names by many voice assistants and conversational agents.

Our contributions are as follows:

1. We investigate the performance of state-of-the-art (SOTA) 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[4].

2. Related work

Developing ASR systems for low-resource languages remains challenging due to the scarcity of training data and resources. As a result, models trained on high-resource languages, such as English, do not perform well on low-resource languages [5]. To address this, researchers have proposed several solutions such as cross-lingual representations where the system learns a shared representation for multiple languages [6], data augmentation techniques [7], and fine-tuning ASR model trained on high-resource languages on low-resource languages [8]. Recent SOTA multilingual ASR models such as Whisper [9] – trained on over 680K hours labeled speech samples, including multilingual speech corpora such as Common Voice [10] – have significantly improved the ASR landscape, outperforming their monolingual counterparts such as HuBERT [11], wavLM [12], and wav2vec2 [13] in various downstream tasks. Despite these

† Equal contribution.

[1] https://techxplore.com/news/2022-09-effective-automatic-speech-recognition.html

[2] Fela is one of Africa's most legendary artists

[3] To "butcher" a name means to mispronounce it, resulting in a significant deviation from the correct pronunciation.

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

10.21437/Interspeech.2023-2122

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.
aws	if you did good has been living at kaduna with his wife, she america or raji mo to
w2v2-lg-960h-lv60-self	fia digo has been living at cadna with his wife shi maca orajimo to truco o
w22-lg-xlsr-53-en	ifia digu has been living at kaduna with his wife shiamaka orajimo tutruku
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 orahjimu tochukwu

breakthroughs, both open source and commercial ASR systems still exhibit racial bias [14], higher error on accented speech [15], and incorrect transcriptions of named entities. Previous studies have highlighted challenges with named entity recognition (NER) for ASR and have investigated various methods to improve NER performance. For instance, French researchers [16] outlined steps for assessing NER in french transcripts of radio broadcasts, while [17] evaluated Chinese accent ASR on an automatic speech query service (AVQS), highlighting the severe limitations of such systems for Mandarin users with multiple accents. More recently, [18, 19] attempted to extract semantic information directly from speech signals using a single end-to-end model that learns ASR and NER tasks together. However, none of this work focuses on named entities in African datasets, which presents a new area of research and its unique challenges.

3. Methodology

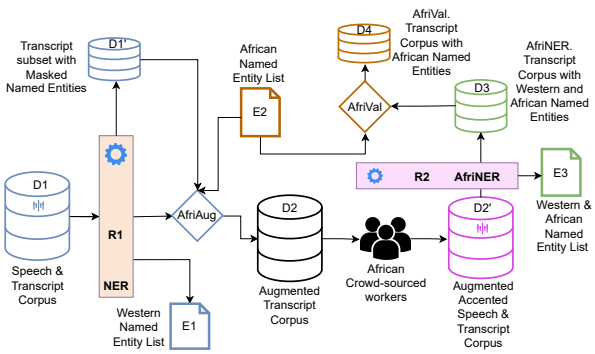

Figure 1: *AfriNames dataset augmentation process.*

3.1. African Named-Entity Augmentation Workflow

Western vs African-named entities: We use the term "Western named entities" to refer to names that are commonly used in Western cultures and languages, such as Laura and Buenos Aires, and that may not have direct translations in African languages [5]. In contrast, we use the term "African named entities" to denote names, places, and cultural references that are derived from African languages and cultures, and that may not be commonly used or recognized outside of those contexts.

Approach: We address the generalization problem as a domain shift, depicted in Figure 1. Our initial dataset, denoted as D_1, consists of Western audio samples X^{E_1} and their corresponding transcripts Y^{E_1}. We employ a pre-trained named entity recognition (NER) model R_1 to extract named entities (NEs) from Y^{E_1}, resulting in the predominantly Western named entity list E_1. To inject African named entities, we mask tokens in randomly selected samples from Y^{E_1} that match the entities in E_1. This process generates the modified dataset D_1' with modified transcripts Y'^{E_1}. We then randomly insert tokens from a curated African named-entity list E_2 to replace the masked tokens in Y'^{E_1}, creating an augmented dataset D_2 with modified transcripts Y^{E_2}. These transcripts are sent to African crowd-sourced workers for recording, resulting in a new corpus named D_2' with augmented pairs $\{(X^{E_2}, Y^{E_2})\}$. This novel dataset comprises accented audio samples and augmented transcript pairs, combining distributions from D_1 and D_2 with Anglo-centric named entities E_1 and African named entities E_2. Next, we use a specialized NER model R_2 to annotate all western and African named entities (called E_3) present in D_2. Using these NER annotations, we select the subset of D_2 with NEs. This NE subset D_3 (called AfriNER) contains accented speech X^{E_3} and corresponding transcripts Y^{E_3} with named entities extracted from both Y^{E_1} and Y^{E_2}. Additionally, using curated African NE list E_2, we also filter Y^{E_3} to create D_4 confirmed to contain African NEs (called AfriVal).

3.2. Datasets

In this study, we primarily explore the AfriSpeech-200 dataset, a 200.91 hours novel accented English speech corpus rich with African-named entities, curated for clinical and general domain ASR using the augmentation process described above. 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 (Table 2).

3.3. AfroAug: African Named-Entity Augmentation

To increase the representation of African named entities, we start with a corpus D_1 using large open-source predominantly western corpora: Wikitext-103 [20] and scrape African entertainment and news websites to increase the representation of African content. We augment this dataset using two main strategies. We curate a list E_2 of approximately 100k African names using a database of 90,000 African names from [21], 965 Nigerian Igbo names from [22], and 1,000 African names obtained

[5] Due to the influences of colonization and globalization, many Western names have been adopted in African cultures. Therefore, while these names may not have direct translations in African languages, they can still be used and recognized in African contexts. Our work focus specifically on African named entities that are derived from African languages.

Table 2: *AfriSpeech-200 Dataset statistics*

	Train	Dev	Test
Duration (hrs)	173.4	8.74	18.77
# General domain clips	21682	1407	2723
Unique Speakers	1466	247	750
Accents	71	45	108
Named Entities Category Counts			
PER	11011	669	1064
ORG	6322	372	279
LOC	3194	192	526

from freely available textbooks, online baby name websites, oral interviews, published articles, and online forums like Instagram and Twitter; and African cities list from Wikipedia [6]. We augment D_1' in three key steps:

1. **Named-Entity Extraction with NER Models:** We leverage off-the-shelf pre-trained NER models [23] and annotate all named-entities in corpus Y^{E_1} to extract the list E_1, tokens tagged with [PER], [LOC], or [ORG]. We mask these tokens $e_i \in E_1$ for a randomly sampled subset of transcripts.
2. **Template Selection:** We manually review, select and validate 140 of these sentences where the replacement of masked tokens with African named entities sounds natural and retains meaning in context. These curated sentences with masked tokens are selected as final templates.
3. **Named Entity Replacement:** We randomly (uniformly) replace all [LOC] tags with African cities from E_2, and all [PER] and [ORG] tags with African names from E_2. We repeat this process 200 times to create text corpus Y^{E_2} consisting of 28,000 novel augmented transcripts combined with transcripts from Y^{E_1} (100,000+ sentences). Y^{E_2} is recorded by crowd-sourced workers. We sample a subset of users from train/dev/test splits for this work.

A real-world example of D^{E_1} is LibriSpeech [24], a 1,000-hours speech-text dataset from English-only audiobooks. The resulting ASR model $M_1^{E_1}$, such as Wav2vec2 [13], therefore, generalizes poorly to African named entities E_2 (Table 1). The pretrained ASR model $M_1^{E_1}$ is thus fine-tuned on the new augmented training dataset D_2', and learns a new mapping $f : X^{E_2} \longrightarrow Y^{E_2}$ resulting in a more robust model $M_1^{E_2}$, adapted to the target distribution D^{E_2}.

4. Experiments

4.1. Benchmarks

We compare SOTA open-source pre-trained ASR models: Whisper [9], Wav2vec2 [13], XLSR [25], Hubert [11], and WavLM [12], with commercial ASR systems. We refer readers to the respective papers for details on pre-training corpora, model architecture, and hyperparameters. We compare 4 model categories: (1) **Monolingual Models** pre-trained or fine-tuned exclusively on predominantly western transcripts, western English speech, and western named-entities (2) **Multilingual Models** pre-trained on transcripts from multiple domains, western and accented speech, but with minimal amounts of African named-entities (3) **Commercial ASR APIs** (4) **Ours** finetuned on western and African-named entities paired with audios in accented African English.

[6] https://en.wikipedia.org/wiki/List_of_cities_in_Africa_by_population

4.2. Fine-tuning

We select two best-performing open-source models from section 4.1 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 [25]: 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 [9]: 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 [26], AdamW [27], 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 24-48 hrs.

4.3. Evaluation

Word Error Rate (WER) and Character Error Rate (CER) are common metrics for evaluating ASR models. WER measures word errors, CER measures character errors. Lower values are better for both.

4.3.1. AfriNER: Named-Entity Evaluation

To evaluate ASR performance on named entities (NEs), we need a reliable way to identify samples in Y_2 with NEs. Ground truth transcripts Y_2 contain E_1 and E_2 entities, jointly called E_3. To extract all samples in Y_2 with NEs in E_3, we run NER inference on all test samples in Y_2 using a specialized performant NER model R_2 [7] from [31] that jointly predicts the set of African and western named entities E_3. We select 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. R_2 is also able to identify unknown African named entities in Y_2 not sourced from E_2 (but present in Y_1). We denote this subset Y_3 (Afri-NER). For each model, we compute WER on corresponding model predictions Z_3.

4.3.2. Sentence-level AfriValidation: African Named-Entity Validation

Our primary goal is to evaluate $M_1^{E_1}$s and $M_1^{E_2}$ on transcripts with "African" NEs. To isolate samples with African NEs, we extract the subset of Y_2 from the test partition with any NEs from E_2 to create the AfriVal subset. Because these sentences are known to contain African NEs, they are Afri-Validated, and guarantee we can reliably evaluate ASR models on predicted transcripts with African NEs. For each model, we compute WER on corresponding model predictions.

4.3.3. Character-level AfriValidation

Since sentence-level WER is impacted by non-NE tokens, we compute CER on NE tokens by isolating them as follows: 1) We run R_2 on model predicted transcript Z_2 and Z_3 to obtain

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

Table 3: *WER results on Afrispeech test samples.* ***All*** *is mean WER across all test samples.* ***No-NER*** *is mean WER across samples with NO predicted named entities (NEs).* ***AfriNER*** *is mean WER across all sentences WITH predicted NEs.* ***AfriVal*** *is mean WER across AfriValidated samples.* ***char-AfriNER*** *and* ***char-AfriVal*** *are mean CER on AfriNER and AfriVal respectively.* ***char-AfriNER*** *and* ***char-AfriVal*** *concatenates the NEs in the predicted and reference transcripts.*

Model	Params	Training or Finetuning data	WER				CER	
			All (#2364)	No-NER (#1029)	AfriNER (#971)	AfriVal (#229)	char-AfriNER	char-AfriVal
Baseline								
wav2vec2-large-960h	317M	Monolingual	0.641	0.565	0.696	0.802	0.861	0.986
Monolingual Fine-tuning: Open-Source SOTA pre-trained Models						0.718 Monolingual Mean WER		
wav2vec2-large-960h-lv60-self	317M	Monolingual	0.533	0.458	0.584	0.683	0.808	0.978
hubert-xlarge-ls960-ft	317M	Monolingual	0.562	0.487	0.613	0.701	0.803	0.986
wavlm-libri-clean-100h-large	317M	Monolingual	0.631	0.562	0.680	0.769	0.864	0.984
Multilingual Fine-tuning: Open-Source SOTA pre-trained Models						0.506 Multilingual Mean WER		
whisper-large	1550M	Multilingual	0.240	0.187	0.300	0.412	**0.565**	0.855
whisper-medium	769M	Multilingual	0.276	0.206	0.352	0.488	0.607	0.913
wav2vec2-large-xlsr-53-english	317M	Multilingual	0.506	0.447	0.550	0.617	0.772	0.965
Commercial ASR APIs						0.588 Commercial Mean WER		
Azure[28]	-	-	0.340	0.273	0.402	0.509	0.674	0.946
GCP[29]	-	-	0.534	0.464	0.603	0.700	0.827	0.991
AWS[30]	-	-	0.354	0.279	0.426	0.556	0.735	0.970
AfriSpeech Finetuning (Ours)						0.160 AfriSpeech Mean WER		
whisper-medium-AfriSpeech	769M	Monolingual, AfriSpeech	**0.186**	**0.172**	**0.198**	**0.108**	0.576	**0.704**
xlsr-53-english-AfriSpeech	317M	Monolingual, AfriSpeech	0.236	0.211	0.258	0.212	0.622	0.816

predicted NE tokens with > 0.8 confidence score. To mitigate the impact of NER errors from R_2, for each ground truth and predicted sentence, we concatenate all NER tokens $e_i \in E_3$ from Y_3 and all $z_i \in Z_3$ removing all spaces and compute CER. For selected pre-trained and commercial ASR models $M_1^{E_1}$, as well as fine-tuned models $M_1^{E_2}$, we evaluate WER and CER on samples containing one or more named entities and present single run results in Table 3.

5. Results and Discussion

5.1. African named entities are challenging

The baseline model in Table 3 demonstrates the dominant trend in our results. WER on all samples (column 4, All) improves by 13.6% (relative) when samples with named entities are EXCLUDED (column 5, No-NER), worsens by 11.7% (relative) when samples with named entities (western + African) are isolated (column 6, AfriNER). Performance sinks by 29.7% (relative) on the subset of Afrivalidated examples (column 7, AfriVal)– samples with African-named entities from $E2$. This pattern is consistent across all model categories except Ours where we observed a 41.9% (whisper) and 10.2% (xlsr) relative WER improvement on AfriVal sentences.

5.2. Training data bias

As shown in Table 3, multilingual/multitask pre-training outperforms monolingual pre-training/fine-tuning. Multilingual/multitask models [9, 25, 32] learn more useful representations, are more linguistically diverse, robust, and generalize better to accented speech when compared with monolingual models fine-tuned on datasets (e.g. Librispeech [24] and Switchboard [33]) with predominantly western NEs and western accents. After fine-tuning on AfriSpeech with African NEs and accented speech, our best model, whisper-medium improves on the baseline by 81.5% compared to 16.4% for the pre-trained model.

5.3. Multilingual pretraining is insufficient

Despite extensive pretraining on 680k hours of multilingual data (90 languages), the fine-tuned model outperforms the pre-trained model by 77.9% (relative). Our results demonstrate that multilingual/multi-task pretraining is inadequate as these SOTA models make several mistakes with African-named entities. Fine-tuning results show that our approach is effective in mitigating bias in these large models.

5.4. Character-Level analysis

When named entities are isolated as described in Section 4.3.2, we observe that our fine-tuned whisper-medium model worsens by 1.9% (relative) in comparison to the pre-trained whisper-large model (column 8, char-AfriNER). This may be due to the significantly higher number of parameters in whisper-large generalizing better to certain named entities. However, when evaluated on the Afrivalidated dataset (column 9, char-AfriVal), our fine-tuned whisper-medium model outperforms both pre-trained whisper-large and medium models (relative gain of 17.7%, and 22.9% respectively). These results further support our claim that the presence of African-named entities is crucial for achieving better performance in ASR models.

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 rescore the raw ASR transcript. This is especially destructive for African-named entities. Because these named entities (e.g."Ifeadigo") are missing from LM training data, where the probability of sequences with African NEs 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, shifting the distribution for robustness in the African context.

7. References

[1] I. 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*, 2021.

[2] T. Desot, F. Portet, and M. Vacher, "Towards end-to-end spoken intent recognition in smart home," in *2019 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*. IEEE, 2019.

[3] D. I. Adelani, J. Abbott, G. Neubig, D. D'souza, J. Kreutzer, C. Lignos, C. Palen-Michel, H. Buzaaba, S. Rijhwani, S. Ruder *et al.*, "Masakhaner: Named entity recognition for african languages," *Transactions of the Association for Computational Linguistics*, 2021.

[4] M. Nguyen and Z. Yu, "Improving named entity recognition in spoken dialog systems by context and speech pattern modeling," 2021.

[5] Ł. Lepak, K. Radzikowski, R. Nowak, and K. J. Piczak, "Generalisation gap of keyword spotters in a cross-speaker low-resource scenario," *Sensors*, 2021.

[6] G. Lample, A. Conneau, M. Ranzato, L. Denoyer, and H. Jégou, "Word translation without parallel data," in *International Conference on Learning Representations*, 2018.

[7] S. Y. Feng, V. Gangal, J. Wei, S. Chandar, S. Vosoughi, T. Mitamura, and E. Hovy, "A survey of data augmentation approaches for nlp," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021.

[8] A. Anaby-Tavor, B. Carmeli, E. Goldbraich, A. Kantor, G. Kour, S. Shlomov, N. Tepper, and N. Zwerdling, "Do not have enough data? deep learning to the rescue!" in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020.

[9] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *arXiv preprint arXiv:2212.04356*, 2022.

[10] R. Ardila, M. Branson, K. Davis, M. Kohler, J. Meyer, M. Henretty, R. Morais, L. Saunders, F. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, 2020.

[11] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2021.

[12] S. Chen, C. Wang, Z. Chen, Y. Wu, S. Liu, Z. Chen, J. Li, N. Kanda, T. Yoshioka, X. Xiao *et al.*, "Wavlm: Large-scale self-supervised pre-training for full stack speech processing," *IEEE Journal of Selected Topics in Signal Processing*, 2022.

[13] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, 2020.

[14] A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Toups, J. R. Rickford, D. Jurafsky, and S. Goel, "Racial disparities in automated speech recognition," *Proceedings of the National Academy of Sciences*, 2020.

[15] A. Hinsvark, N. Delworth, M. Del Rio, Q. McNamara, J. Dong, R. Westerman, M. Huang, J. Palakapilly, J. Drexler, I. Pirkin *et al.*, "Accented speech recognition: A survey," *arXiv preprint arXiv:2104.10747*, 2021.

[16] S. Galliano, G. Gravier, and L. 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.

[17] K. Xiao and Z. Qian, "Automatic voice query service for multi-accented mandarin speech," in *2021 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. IEEE, 2021.

[18] S. Mdhaffar, J. Duret, T. Parcollet, and Y. Estève, "End-to-end model for named entity recognition from speech without paired training data," in *Interspeech 2022*, 2022.

[19] A. Caubrière, S. Rosset, Y. Estève, A. Laurent, and E. Morin, "Where are we in named entity recognition from speech?" in *Proceedings of the 12th Language Resources and Evaluation Conference*, 2020.

[20] S. Merity, C. Xiong, J. Bradbury, and R. Socher, "Pointer sentinel mixture models," in *International Conference on Learning Representations*, 2017.

[21] R. Anderson, A. Borucki, D. D. Da Silva, D. Eltis, P. Lachance, P. Misevich, and O. 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*, 2013.

[22] H. I. Okagbue, A. A. Opanuga, M. O. Adamu, P. O. Ugwoke, E. C. Obasi, and G. A. Eze, "Personal name in igbo culture: a dataset on randomly selected personal names and their statistical analysis," *Data in brief*, 2017.

[23] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020.

[24] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015.

[25] J. Grosman, "Fine-tuned XLSR-53 large model for speech recognition in English," https://huggingface.co/jonatasgrosman/wav2vec2-large-xlsr-53-english, 2021.

[26] P. Micikevicius, S. Narang, J. Alben, G. Diamos, E. Elsen, D. Garcia, B. Ginsburg, M. Houston, O. Kuchaiev, G. Venkatesh *et al.*, "Mixed precision training," in *International Conference on Learning Representations*, 2018.

[27] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations*, 2019.

[28] Microsoft, "Real-time speech-to-text," https://speech.microsoft.com/portal/speechtotexttool, 2023, [Online; accessed 1-December-2022].

[29] Google, "Speech-to-text," https://cloud.google.com/speech-to-text/, 2023, [Online; accessed 1-December-2022].

[30] Amazon, "Amazon transcribe," https://aws.amazon.com/transcribe/, 2023, [Online; accessed 1-December-2022].

[31] D. I. Adelani, G. Neubig, S. Ruder, S. Rijhwani, M. Beukman, C. Palen-Michel, C. Lignos, J. O. Alabi, S. H. Muhammad, P. Nabende, C. M. B. Dione, A. Bukula, R. Mabuya, B. F. P. Dossou, B. K. Sibanda, H. Buzaaba, J. Mukiibi, G. Kalipe, D. Mbaye, A. Taylor, F. Kabore, C. C. Emezue, A. Aremu, P. Ogayo, C. W. Gitau, E. Munkoh-Buabeng, V. M. Koagne, A. A. Tapo, T. Macucwa, V. Marivate, E. Mboning, T. R. Gwadabe, T. P. Adewumi, O. Ahia, J. Nakatumba-Nabende, N. L. Mokono, I. M. Ezeani, C. I. Chukwuneke, M. Adeyemi, G. Hacheme, I. Abdulmumin, O. Ogundepo, O. Yousuf, T. M. Ngoli, and D. Klakow, "Masakhaner 2.0: Africa-centric transfer learning for named entity recognition," *ArXiv*, 2022.

[32] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu *et al.*, "Conformer: Convolution-augmented transformer for speech recognition," 2020.

[33] J. J. Godfrey, E. C. Holliman, and J. McDaniel, "Switchboard: Telephone speech corpus for research and development," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on*. IEEE Computer Society, 1992.