On the Tolerance of Repetition Before Performance Degradation in Kiswahili Automatic Speech Recognition

Kathleen Siminyu† **Kathy Reid**‡ **Rebecca Ryakitimbo**#
Britone Mwasaru⋆ **Chenai Chair**⋆
†The Distributed AI Research Institute
‡School of Cybernetics, Australia National University
# Kuza STEAM Generation ⋆ Independent

Abstract

State of the art end-to-end automatic speech recognition (ASR) models require large speech datasets for training. The Mozilla Common Voice project crowd-sources read speech to address this need. However, this approach often results in many audio utterances being recorded for each written sentence.

Using Kiswahili speech data, this paper first explores how much audio repetition in utterances is permissible in a training set before model degradation occurs, then examines the extent to which audio augmentation techniques can be employed to increase the diversity of speech characteristics and improve accuracy.

We find that repetition up to a ratio of 1 sentence to 8 audio recordings improves performance, but performance degrades at a ratio of 1:16. We also find small improvements from frequency mask, time mask and tempo augmentation. Our findings provide guidance on training set construction for ASR practitioners, particularly those working in under-served languages.[1]

1 Introduction

Automatic Speech Recognition(ASR) is the process of converting acoustic speech into text (Washani and Sharma, 2015). This task has gained significance with the increased use of computing systems by humans via voice commands. End-to-end (E2E) speech recognition models have several components that contribute to the development of the overall system. These include an acoustic model which gives the most likely acoustic unit (phone) based on the acoustic properties of the input signal, a language model which can represent the linguistic form of a language, and thus defines the words in this language and how likely they are to occur together and a lexicon which explains the vocabulary at the phone-level (Leino et al., 2015). These models require large volumes of speech data for training.

The accuracy of E2E ASR models is typically evaluated using two metrics - word error rate (WER) and character error rate (CER). WER and CER are defined as the number of word or character insertions, omissions and substitutions in a transcription, divided by the number of matching words or characters respectively (Kamath et al., 2019). WER measures the accuracy of the language model while CER measures the accuracy of the acoustic model. We acknowledge that the suitability of these metrics is contested per Aksënova et al. (2021).

In a bid to reduce error rates, the ASR community continues to call for greater quantities of data to train systems, going from 50 to 500 to 500 hours of speech (Moore, 2003). Speech Recognition datasets are composed of recordings of speech which are accompanied by corresponding texts or transcripts. They can be obtained by taking existing audio recordings, having them transcribed, splitting them into shorter audio segments and aligning the recordings to their transcriptions. This process describes the creation of a spontaneous speech dataset, which is speech produced by a speaker in an informal, dynamic, unrehearsed, casual manner (Tucker and Mukai, 2023). Datasets such as the FAU Aibo Emotion Corpus (Batliner et al., 2008) contain spontaneous speech. In some cases, the script or transcription comes first then audio recordings are created through speakers being prompted to read out the script while recording themselves, resulting in an elicited or read speech dataset. Datasets such as Multilingual LibriSpeech (Pratap et al., 2020) and Mozilla Common Voice (Ardila et al., 2019) are examples of read speech datasets. The Mozilla Common Voice(MCV) dataset is a multilingual speech corpus developed for Auto-

[1] This research was conducted while the authors - Kathleen Siminyu, Rebecca Ryakitimbo, Britone Mwasaru and Chenai Chair - were affiliated to Mozilla Foundation.

matic Speech Recognition purposes (Ardila et al., 2019). The data collection efforts are entirely crowd-sourced through organising and engaging language communities.

This paper documents work that has focused on the Kiswahili language dataset available on MCV. Kiswahili is a Bantu language originally spoken by the Swahili people of Eastern Africa. It is one of the official languages of the East African Community in addition to being a national language in Tanzania, Kenya, the Democratic Republic of Congo and Uganda. Kiswahili has over 200 million speakers[2]. It is the most widely spoken African language.

The efforts in building the Kiswahili dataset on MCV, are described in greater detail in §3.1. This dataset contains an underlying text corpus of 134,653 Kiswahili sentences and from this, over 700,000 audio clips have been recorded totalling 1,081 hours of audio data. There are over 1,454 individual speakers that have contributed their voices to create the dataset. While these efforts are commendable, the resulting dataset for Kiswahili, MCV 16[3] in some cases has up to 16 corresponding audio recordings to a single sentence. These are instances of different speakers having recorded themselves reading the same sentence out loud.

This work is to help us determine how best to utilise our dataset in training a neural model for speech recognition that is able to generalise well. This set of experiments examines: 1) how much audio repetition, in relation to a sentence, can be included in a training dataset, before this leads to a degradation of performance of the output model, and 2) whether audio augmentation techniques can be employed to reduce repetition and increase diversity (speaking rate, background noise and interference, pitch) within our dataset.

We find that repetition up to a ratio of 1 sentence to 8 audio recordings improves performance, but performance degrades at a ratio of 1:16. Additionally, various augmentation techniques lead to improvements; time mask augmentation led to an improvement of up to 4.2%, tempo augmentation led to an increase of up to 3.36% and frequency mask augmentation led to an increase of up to 2.4%.

2 Prior Work

We infer from speech recognition literature, specifically from the descriptions of the creation of elicited speech datasets, which are comparable to MCV, such as Librispeech (Panayotov et al., 2015) and Multilingual Librispeech (Pratap et al., 2020), that in an ideal data setting we expect a 1:1 ratio of audio to transcript to ensure adequate variety of content in the dataset. In these datasets, the data is derived from read audio books and each book contains only one accompanying audio recording.

While machine learning literature suggests that the more data available to train a model, the better an output system would be (Halevy et al., 2009; Brill, 2003), there is also literature indicating that, particularly in supervised learning scenarios, insufficient samples for learning or repetition within a dataset would lead a model to overfit during training (Ying, 2019). Overfitting is an issue in supervised machine learning where a model is unable to generalize on unseen data, thus performing poorly, despite appearing to generalise on observed data available in the training set (Russell and Norvig, 2010).

Augmentation of data is another strategy that can be used to prevent overfitting. Data augmentation is the generation of synthetic data from already existing data (Ko et al., 2015). Rebai et al. (2017) have shown that data augmentation techniques can be employed to instances where limited data is available with the intent to modify instances of the data so as to increase the amount of training data. These techniques are also used to improve the performance of resulting systems as they serve to introduce variety in training data; frequency and pitch masks can help make the model more robust when faced with background noise and interference in audio recordings, pitch and tempo augmentation can serve to add more 'speakers' to a dataset by adding speakers with the same articulation patterns or creating speakers with new articulation patterns, respectively (Zhang et al., 2023; Zevallos et al., 2022; Ying, 2019).

3 Methodology

3.1 Data Collection

There are several stages in the data collection process on MCV.

3.1.1 Creation and Collection of Sentences

Existing texts can be added onto the platform provided they are in the public domain. This is a requirement because the entirety of the MCV dataset is licensed as CC Zero (CC0). This is a Creative

[2] Swahili gaining popularity globally

[3] Mozilla Common Voice Kiswahili dataset

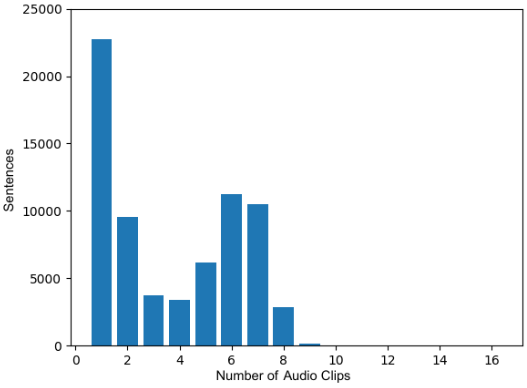

Figure 1: The y-axis shows the number of number of sentences and the x-axis shows the accompanying audios recorded for instances in the validated set.

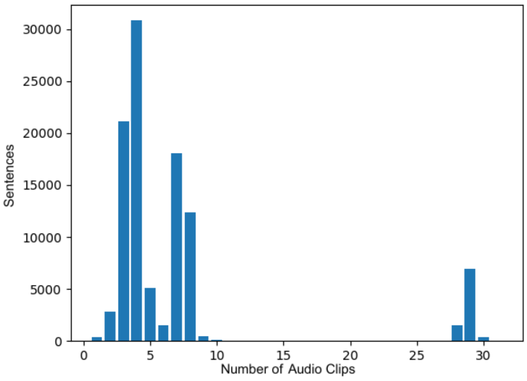

Figure 2: Similar to figure 1, the y-axis shows the number of sentences and the x-axis shows the accompanying audios recorded for sentences. In this case, in addition to the number of clips that have been validated, we include those that have been invalidated and those that are yet to be validated.

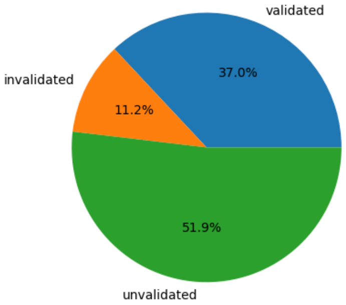

Figure 3: The pie chart shows the percentage of audios that have either been validated, invalidated - has received two or more down-votes - or are yet to be validated (i.e. unvalidated) in version 16 of the MCV Kiswahili dataset.

Commons License that allows creators to give up their copyright and put their works into the worldwide public domain. CC0 allows re-users to distribute, remix, adapt and build upon the material in any medium or format, with no conditions [4]. Where there is existing text with ownership attributed to an individual or an organisation, and they are willing to waive these rights so that the content is added onto the platform, they need to sign a waiver giving MCV permission [5]. In our work, community initiatives and events have been organised in support of the creation of original Kiswahili texts for addition onto the platform. One example is a partnership with Hekaya Arts Initiative, a writers collective based in Mombasa Kenya, which saw us organise a series of writing competitions. Submission to the competitions have been added onto the MCV platform and winners in each edition were awarded prizes [6].

3.1.2 Validation of Sentences

These sentences then need to go through a validation process. Each language may have slightly differing requirements for sentence validation. Some general reviewing criteria include ensuring that the spelling and grammar in a sentence are correct, that they are natural and conversational (should be easy to read the sentence), that there is no use of abbreviations or acronyms and that there are no digits in the source text. These should all be written down in full and in text format to avoid ambiguity when reading aloud. Each sentence requires at least 2 up votes to be added onto the MCV platform for voice contributions.

3.1.3 Collection of audio recordings

Contributors can choose the 'Speak' feature on the MCV platform, where a single sentence at a time is displayed and the contributor is prompted to record themselves reading it out loud. New sentences, those that do not yet have an accompanying recording, are prioritised on the platform. Once each sentence has at least one accompanying recording, the sentences then begin being looped over again. Should the voice contributions come in faster than the text contributions, some sentences will have more than one accompanying audio. The platform

[4] Creative Commons Licenses

[5] Common Voice Contribution Agreement for Pre-existing Works

[6] Common Voice, Hekaya Arts Initiative Announce Kiswahili Writing Competition Winners

tracks what sentences an individual contributor has already recorded an audio for, provided they are signed in. They will therefore ideally contribute only one audio recording per sentence. In the event that a 'super' contributor provides an audio for the entire underlying text corpus, they will then start to be presented with sentences for speech elicitation that they have already recorded an audio for. It is therefore possible for a single sentence to have more than one accompanying audio, provided by different speakers. It is also possible, however rare, for there to be duplicates of an individual speaker contributing more than one audio to a single sentence. These features of the platform impact characteristics of the dataset, in terms of creating opportunity for repetition. In our work collecting data for Kiswahili, we experienced significant challenges in accessing text data early on in the project. Our efforts in collecting text were happening concurrently with community efforts to contribute and validate audios. We soon found ourselves with up to 30 audios per sentence for the texts that were seeded onto the platform in the very beginning as evidenced by figure 2, which shows the amount of repetition in all available sentences and their accompanying audios and figure 1, which shows the amount of repetition in the validated subset of the dataset.

3.1.4 Validation of audio recordings

Contributors can choose the 'Listen' feature where they will be prompted to play already recorded audios and to validate them, by giving them a thumbs up, or invalidate them, by giving them a thumbs down, depending on whether or not the audios fit the reviewing criteria provided. This includes listening to ensure that the contents of the audio align with the accompanying text, that the speaker is audible enough and that they do not hesitate or stammer. This validation is important when producing a speech recognition corpus as they will directly impact the quality of models produced. If the transcript and audio recording are not accurate, then the model becomes less likely to be accurate. This validation work is crowd-sourced and takes place in community events[7]. The participants are therefore not trained linguists or language professionals for the most part, they are native speakers of the language. Given the diversity of Kiswahili speakers, and the varying accents available, it is important to curate a diverse group of validators to ensure that this diversity is maintained in the dataset that is curated for Machine Learning. We found that voice validation is not as popular as voice contributions in our community activities. This has resulted in only 48.2% of our dataset having undergone validation (37% validated and 11.2% invalidated), as of the MCV 16 release. This is approximately 400 hours of data that is considered fit for use, i.e. validated, compared to the 1081 hours available. Due to concerns about the quality of unvalidated data, we use only data that is validated and more than half of the data available is left unutilised. Figure 3 shows the validation rate of Kiswahili data in the MCV 16 release.

3.2 Data Pre-processing

The Kiswahili dataset on MCV 16 comes with seven files:

- `validated.tsv` - contains information on the audios that have been validated, i.e. have received at least 2 up-votes
- `reported.tsv` - contains information about sentences that have been reported to have a grammatical or spelling error, having offensive language, having a different language or being difficult to pronounce
- `invalidated.tsv` - contains information of audios that have received at least 2 down-votes, and less than two up-votes
- `other.tsv` - contains information of audios that have neither been validated nor invalidated
- `train.tsv` - contains the list of audios included as part of the training data
- `dev.tsv` - contains the list of audios included as part of the development data used to validate the model's learning during training
- `test.tsv` - contains the list of audios included as part of the test data

In this work, we curate our own experimental splits as opposed to using the train, dev and test splits provided with MCV. We make the decision to use only the instances that have been validated, i.e., those that are listed in the `validated.tsv` file as having been reviewed and verified by Kiswahili

[7] Common Voice Kiswahili Festival Brings Community Together To Grow Dataset

speakers. There are approximately 400 hours of validated data in the Kiswahili dataset.

We filter out several subsets that have been created specifically for purposes of evaluation of certain demographic groups. These are data for dialects and variants that are closely related to Kiswahili: Kiunguja, Kibajuni, Kimakunduchi, Kimvita, Kipemba, Kitumbatu and Kiswahili cha Bara ya Tanzania (Kiswahili from Inland Tanzania). These subsets were developed through working with linguists and language experts, work that has been documented (Siminyu et al., 2022).

We investigated the existing CorporaCreator repository[8], a command line tool to create Mozilla Common Voice corpora for use in this work, however it did not provide the flexibility to curate our own evaluation sets, particularly how big they should be. We found that while it is useful to be able to select the number of audio repetitions included in the training set, this changed the composition of the development and test sets in each instance, a behaviour which makes the first set of experiments in this work incomparable. We therefore chose to create our own scripts for data preprocessing.

We split our data into 3 sets, a training set, a development set and a test set, in the ratio 60:20:20.

In constituting our training, development and test sets, we consider several factors:

- That all audios corresponding to a single sentence should only appear in one set
- That all audios contributed by a single speaker should also all be in only one set
- Where a single speaker may have contributed to an individual sentence more than once, we drop duplicate instances

4 Experiments

4.1 Experiment 1: More Data versus Less Repetition Trade-off

In the first set of experiments, we consider a trade-off in constituting the training set. On one hand, an increase in audio repetition, in relation to a single sentence, creates more training data. On the other hand, repetition of audio recordings relative to an individual sentence may decrease the performance of the output model due to overfitting. We increase the text to audio ratio following a geometric progression up to the maximum number of repetitions available, in this case 16, with the intent of drawing a curve that can visualize the results and determine whether there is a point at which more data and more repetition leads to a degradation of performance in the output models. The data is constituted in the following settings;

- 1:1 - each sentence with 1 accompanying audio recording
- 1:2 - each sentence with 2 accompanying audio recordings
- 1:4 - each sentence with 4 accompanying audio recordings
- 1:8 - each sentence with 8 accompanying audio recordings
- 1:16 - each sentence with 16 accompanying audio recordings

We use the Coqui AI Speech-to-Text(STT) toolkit[9] for these experiments. The Coqui STT architecture consists of a recurrent neural network(RNN) with 5 hidden layers, where the first three and the fifth layers are non-recurrent and use a clipped rectified-linear (ReLu) activation function while the fourth layer is a bidirectional recurrent layer. The CTC loss function is used by the network. The system is integrated with an N-gram language model. To identify the ideal hyper-parameter settings, we run several iterations of the experiment with a 1:1 mapping of the data with the following settings:

- `-n_hidden`: 1024, 2048, 5024
- `-reduce_lr_on_plateau`: true
- `-plateau_epochs`: 10
- `-plateau_reduction`: 0.025, 0.05, 0.1
- `-early_stop`: true
- `-es_epochs`: 25
- `-es_min_delta`: 0.01, 0.02, 0.05
- `-dropout_rate`: 0.3
- `-epochs`: 60 (for our hyper-parameter search)

[8] CorporaCreator Github repository

[9] Coqui STT Github repo

Once we have selected the ideal hyperparameters for our experiment, we do training runs with the different sentence to audio ratio settings; 1:1, 1:2, 1:4, 1:8 and 1:16. We then use the best score as a baseline for our second set of experiments.

4.2 Experiment 2: Data Augmentation for More Data

In this set of experiments, we further wish to explore methods that allow us to make maximum use of the data available to us, in spite of the repetition. We explore the use of audio data synthesis methods to augment subsequent repetitions of audio recordings relative to an individual sentence. The following augmentations are applied to the audio recording repetitions:

- Pitch augmentation shifts the pitch of a waveform by scaling it on the frequency axis. By shifting the pitch, we attempt to add to the variety of "speakers" in the dataset, (Bellettini and Mazzini, 2008) particularly as it relates to age given the skew of available data towards younger speakers (Shahnawazuddin et al., 2020)

- Tempo augmentation changes the playback tempo by scaling the waveform along the time axis. This will help our models become robust to speakers with varying speaking rates (Ko et al., 2015)

- Frequency mask augmentation sets frequency-intervals within the augmented samples to zero (silence) at random frequencies. This helps the model to be robust when it encounters background noise and other interferences in audios (Park et al., 2019)

- Time mask augmentation sets time-intervals within the augmented samples to zero (silence) at random positions. This adds variety in a manner similar to the frequency mask, by making models robust to background noises and interferences (Park et al., 2019)

For this experiment, we use the data from the "1:4" (1 sentence with 4 accompanying audio recordings) setting and the results of the first experiment as our baseline, because this setting represents an acceptable amount of repetition before additional data becomes noisy.

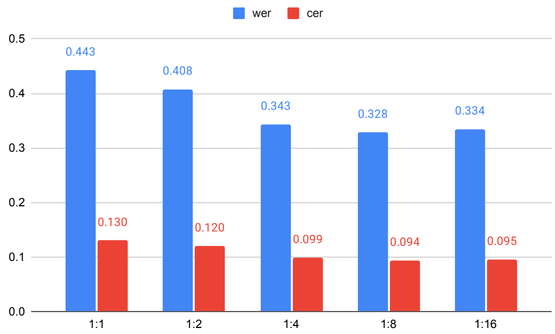

Figure 4: The character error rates and word error rates for the different sentence to audio recordings ratio settings. These steadily decrease with increasing repetition until you reach the 1:8 setting after which there is an increase.

In this experimental setup, given 4 audio recordings relative to one sentence, we vary the ratio of original audio recordings to augmented audio recordings as follows:

- 3:1 - 3 audios in their original form and the final audio recording is augmented. In this case 25% of the data is augmented.

- 2:2 - 2 audios in their original form and 2 of the subsequent repetitions are augmented. In this case 50% of the data is augmented.

- 1:3 - 1 audio is in its original form and 3 of the subsequent repetitions are augmented. In this case, 75% of the data is augmented.

The Coqui AI STT toolkit has implemented a pre-processing pipeline with various augmentation techniques. This feature allows us to set a probability value for each augmentation used. We therefore use the values 0.25 to achieve the "3:1" setting, 0.5 to achieve the "2:2" setting and 0.75 to achieve the "1:3"

We use the ideal hyperparameters selected in experiment 1 to run our experiments.

5 Results

Figure 4 shows results for the first set of experiments in 4.1. It shows WERs and CERs for the different ’sentence to audio ratio’ experimental settings, i.e. 1:1, 1:2, 1:4, 1:8 and 1:16. We see a steady decline in both the WERs and CERs, which is consistent with our expectation that with more data, the models’ overall performance improves. This is true up to the 1:8 setting, when we get to

Table 1: The CER obtained as well as the percentage change (denoted as Δ) given the baseline when 25%, 50% and 75% of the data is augmented using frequency mask, time mask, tempo and pitch augmentations.

	25% augmentation		50% augmentation		75% augmentation	
	CER	Δ	CER	Δ	CER	Δ
frequency mask	0.097	1.12%	0.096	2.40%	0.096	2.23%
time mask	0.094	4.20%	0.099	-0.44%	0.095	2.65%
tempo	0.100	-2.20%	0.095	3.36%	0.099	-1.20%
pitch	0.319	-224.42%	0.101	-3.39%	0.112	-14.57%

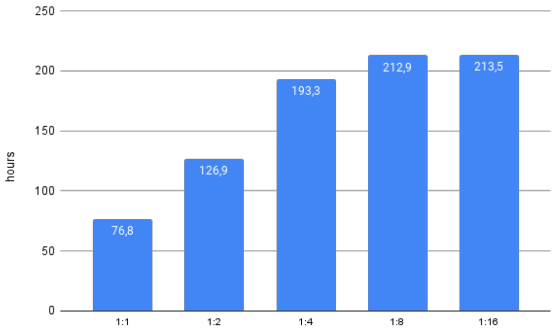

Figure 5: The amount of data, in terms of hours, contained in each of the ’sentence to audio ratio’ settings in experiments in 4.1

the 1:16 setting, the performance is seen to deteriorate. This deterioration in performance may be indicative of how much repetition is acceptable, demonstrating that 16 audio instances of the same sentence is too much despite the additional variability introduced by each subsequent audio being read out by a different speaker. Interrogating the amount of data in each setting reveals that the difference between the 1:4 and 1:8 settings is 19 hours while that between the 1:8 and 1:16 settings is 0.6 hours, as shown in figure 5. In investigating duplicate instances, we found evidence of noise. Audios that were included as duplicates beyond the fourth and eighth instance were likely to have received both up-votes and down-votes from the validation process, e.g. 2 up-votes and 1 down-vote, 4 up-votes and 3 down-votes.

We then used the results of the 1:4 experimental setting as our baseline score for comparison in the second set of experiments to exclude these contentious instances.

The results of the second experiment are shown in Table 1. We find that the frequency mask augmentation consistently led to an improvement of up to 2.4% in CER over the baseline score obtained in the first experiment. This is likely due to the great variation of acoustic settings represented in the data, given that this dataset is crowd sourced by communities. It is likely that frequency mask augmentation leads to zero-ing out of noise in audio samples enabling the model to learn more from the speech to be transcribed.

Time mask augmentation led to an improvement in performance when 25% and 75% of the dataset is augmented, up to 4.2%, but a decrease in performance when 50% of the dataset is augmented. Similar to the frequency mask, the time mask leads to zero-ing out of time steps which possibly contain noise.

Tempo augmentation only led to an improvement when 50% of the dataset is augmented and pitch augmentation did not lead to any increase in performance but showed a shocking decrease of -224.42% when 25% of the dataset is augmented. The frequency axis was scaled by a pitch factor of 0.1 to 0.3, which implies a significant lowering of the pitch far below the original which led to audios becoming low-pitched and likely unintelligible. A better approach would have been to alter the pitch progressively by octave, i.e. 0.5 to take it one octave down, 0.25 to take it two octaves down and 2.0 to raise the pitch by one octave.

There are no consistent gains in perfomance across any of the data augmentation settings(25%, 50% or 75%), leading us to conclude once again that selecting the right augmentation type given the data in question is more pertinent. Overall, we have determined that having more data, despite repetition, can be useful and that the choice of an appropriate data augmentation technique can add greater variation in the dataset making it more useful.

6 Future Work

Given the difficulty faced in identifying Kiswahili text data sources, future work could explore data selection methods for speech utterances and/or text data. This could help eliminate redundancies in data and enable better targeted text data collection, and subsequently speech data collection for low-resource languages. Additionally, as there is a lot of data available that we could not use due to lack of validation, we encourage continued community efforts to validate this data. One limitation of this work is that the experiments have been run on a single dataset and a single language, the work would benefit from scaling up to additional languages and datasets as evidence of generalisability to additional contexts. Finally, given some availability of data for dialects and variants closely related to Kiswahili, it would be great to see how the system developed performs when evaluated on speech from these dialects and variants.

7 Conclusions

In this paper, we articulated the data collection method for read speech in MCV and highlighted the constraint of having many recorded audio utterances from a single written sentence when constructing an ASR training set.

To assess how much sentence repetition is permissible, we trained multiple ASR models on Kiswahili data using varying sentence to audio recording ratios, finding that a ratio of 1:8 is optimal, with performance declining at 1:16. Further, we performed multiple forms of audio augmentation, demonstrating some small improvements in CER for time mask augmentation (4.20% improvement) at 25% augmentation and tempo augmentation (3.36% improvement) at 50% augmentation.

The key take-away for ASR practitioners, particularly those working with under-resourced languages having limited speech data, is that it is worthwhile to include repeated sentences in the training set, however the choice of optimal audio augmentation is likely context-dependent.

Acknowledgements

References

Alëna Aksënova, Daan van Esch, James Flynn, and Pavel Golik. 2021. How might we create better benchmarks for speech recognition? In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 22–34.

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

Anton Batliner, Stefan Steidl, and Elmar Nöth. 2008. Releasing a thoroughly annotated and processed spontaneous emotional database: the fau aibo emotion corpus.

Carlo Bellettini and Gianluca Mazzini. 2008. Reliable automatic recognition for pitch-shifted audio. In *2008 Proceedings of 17th International Conference on Computer Communications and Networks*, pages 1–6. IEEE.

Eric Brill. 2003. Processing natural language without natural language processing. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 360–369. Springer.

Alon Halevy, Peter Norvig, and Fernando Pereira. 2009. The unreasonable effectiveness of data. *IEEE intelligent systems*, 24(2):8–12.

Uday Kamath, John Liu, James Whitaker, Uday Kamath, John Liu, and James Whitaker. 2019. Automatic speech recognition. In *Deep Learning for NLP and Speech Recognition*, pages 369–404. Springer.

Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur. 2015. Audio augmentation for speech recognition. In *Proc. Interspeech 2015*, pages 3586–3589.

Katri Leino et al. 2015. Maximum a posteriori for acoustic model adaptation in automatic speech recognition. Master's thesis.

Roger K Moore. 2003. A comparison of the data requirements of automatic speech recognition systems and human listeners. In *INTERSPEECH*, pages 2581–2584.

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

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779.*

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

Ilyes Rebai, Yessine BenAyed, Walid Mahdi, and Jean-Pierre Lorré. 2017. Improving speech recognition using data augmentation and acoustic model fusion. *Procedia Computer Science*, 112:316–322.

Stuart J Russell and Peter Norvig. 2010. *Artificial intelligence a modern approach.* London.

S Shahnawazuddin, Nagaraj Adiga, Hemant Kumar Kathania, and B Tarun Sai. 2020. Creating speaker independent asr system through prosody modification based data augmentation. *Pattern Recognition Letters*, 131:213–218.

Kathleen Siminyu, Kibibi Mohamed Amran, Abdulrahman Ndegwa Karatu, Mnata Resani, Mwimbi Makobo Junior, Rebecca Ryakitimbo, and Britone Mwasaru. 2022. Corpus development of kiswahili speech recognition test and evaluation sets, preemptively mitigating demographic bias through collaboration with linguists. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 13–19.

Benjamin V Tucker and Yoichi Mukai. 2023. *Spontaneous Speech*. Cambridge University Press.

Nitin Washani and Sandeep Sharma. 2015. Speech recognition system: A review. *International Journal of Computer Applications*, 115(18).

Xue Ying. 2019. An overview of overfitting and its solutions. In *Journal of physics: Conference series*, volume 1168, page 022022. IOP Publishing.

Rodolfo Zevallos, Nuria Bel, Guillermo Cámbara, Mireia Farrús, and Jordi Luque. 2022. Data augmentation for low-resource quechua asr improvement. *arXiv preprint arXiv:2207.06872*.

Yuanyuan Zhang, Aaricia Herygers, Tanvina Patel, Zhengjun Yue, and Odette Scharenborg. 2023. Exploring data augmentation in bias mitigation against non-native-accented speech. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.