ISSN: 2583-9888
DOI: doi.org/10.51219/JAIMLD/amin-jaber/313

URF PUBLISHERS
connect with research world

Journal of Artificial Intelligence, Machine Learning and Data Science

https://urfpublishers.com/journal/artificial-intelligence

Vol: 2 & Iss: 4

Research Article

Enhancement of a Multilingual Model for Automatic Speech Recognition Analysis of Time and Frequency Domain Data Augmentation

Amin Jaber*, Razieh Khamsehashari, Sebastian Möller*

Quality and Usability Lab, Technische Universität Berlin, Germany

Citation: Amin J, Razieh K, Sebastian M. Enhancement of a Multilingual Model for Automatic Speech Recognition Analysis of Time and Frequency Domain Data Augmentation. *J Artif Intell Mach Learn & Data Sci 2024*, 2(4), 1381-1387. DOI: doi.org/10.51219/JAIMLD/amin-jaber/313

Received: 24 September, 2024; **Accepted:** 11 October, 2024; **Published:** 14 October, 2024

***Corresponding author:** Quality and Usability Lab, Technische Universität Berlin, Germany

ABSTRACT

In recent years, significant progress has been made in automatic speech recognition (ASR) systems, especially for languages with abundant transcribed speech data. However, freely accessible models for low-resource languages, particularly those other than English, are still scarce. This study aims to enhance the performance of the prominent multilingual ASR model, Whisper, through fine-tuning and data augmentation techniques. Comprehensive and systematic experiments are conducted on time, frequency and time-frequency domain augmentation strategies across multiple low-resource languages (German, Farsi, Arabic, Russian) to investigate whether these techniques could improve ASR performance in low-resource settings. The study sheds light on the feasibility and limitations of various data augmentations, with a specific emphasis on classic augmentation approaches including audio and spectrogram-based techniques. The findings indicate that audio-based augmentation generally outperforms other evaluated methods. Furthermore, the study explores understanding the effects of data augmentation on model regularization and learning behavior, an aspect that, to our knowledge, has not been extensively explored in ASR. Our findings reveal different behaviors of these techniques not only in the learning process but also in model regularization and generalization. This insight is crucial for integrating these methods into more sophisticated deep learning models, particularly those employing aligned data augmentation in conjunction with classic approaches.

Keywords: Automatic speech recognition; data augmentation; low-resource languages

1. Introduction

TThe development of large-scale speech datasets is often limited by resource constraints, including the availability of labeled data and financial costs. Training ASR models, particularly for low-resource languages, poses additional challenges due to the lack of sufficient labeled samples. Data augmentation techniques provide a promising solution by generating variations of existing audio data, thereby increasing dataset diversity and improving the robustness of ASR systems[1]. Through controlled modifications to audio signals, these techniques can enhance ASR performance across various languages. Recent studies have focused on fine-tuning pre-trained ASR models for low-resource languages, demonstrating the potential of data augmentation to overcome data scarcity[2-4]. However, the impact of specific augmentation techniques on model performance, particularly when data is limited, remains an open question. Gaining a deeper understanding of how data augmentation influences the learning and generalization of ASR models is essential for improving performance across a wide range of languages.

This study investigates the fine-tuning of Whisper[5], a state-of-the-art multilingual ASR model developed by OpenAI, using a range of data augmentation techniques. While Whisper has achieved near-human performance in transcribing English speech[5], its performance in low-resource languages requires further exploration. This work aims to evaluate Whisper's performance across languages with limited data availability by applying time, frequency and time-frequency domain augmentation methods. The study comprehensively compares these augmentation techniques, highlighting their effects on model generalization and regularization. This paper aims to address the following Research Questions (RQ):

- **RQ1:** How do time, frequency and time-frequency domain feature representations as data augmentation techniques affect the generalization ability and representation improvement in ASR models across different low-resoures languages?
- **RQ2:** What are the effects of data augmentation techniques on learning behavior, model regularization and generalization in Whisper and how do these effects vary across languages?
- **RQ3:** How can insights from data augmentation techniques, considering cross-linguistic variations, inform the integration of aligned data augmentation with classic approaches in more sophisticated deep learning-based data augmentation techniques?

The main contributions of this paper are as follows:

- **Data Augmentation Enhancement:** This study conducts a comprehensive analysis of three widely-used data augmentation techniques-time, frequency and time-frequency domain representations. By systematically experimenting with these techniques, the study enhances the generalization ability of ASR models, improving their representation across different linguistic contexts.
- **Novel Insights into ASR:** An in-depth analysis is provided on how data augmentation techniques influence the Whisper ASR model, with a specific focus on learning behavior and model regularization. These aspects have been relatively unexplored in ASR research, particularly for multilingual models like Whisper. The findings highlight the distinct effects of each augmentation technique on both the learning process and generalization, offering valuable insights for integrating these methods into advanced ASR systems.
- **Cross-Linguistic Evaluation:** The study offers a detailed cross-linguistic evaluation of Whisper's performance, analyzing how data augmentation techniques impact generalization across different languages. By focusing not just on a single language but exploring multilingual performance, the research provides valuable insights into how augmentation techniques should be tailored for specific linguistic challenges in ASR.
- **Framework for Future Integration:** The findings offer a framework for integrating classic and aligned data augmentation techniques into more sophisticated deep learning models. This contribution provides practical guidelines for enhancing model regularization and performance in future ASR systems, especially when dealing with low-resource languages.

2. Augmentation Approaches

Data augmentation approaches involve employing techniques intended to expand the size and diversity of training datasets, thereby improving the performance and generalization of ASR models. By leveraging multiple augmentation methods, these approaches contribute significantly to enhancing ASR capabilities. These techniques are typically classified into two main categories. The first is Audio-based Augmentation (AbA), which entails directly altering audio samples[6,7] through various modifications. The second category is spectrogram-based augmentation, which involves adjusting the spectrogram of each sample[8].

A. Audio-based Augmentations

In this study we explore several variations of AbA. The pitch shifting augmentation involves randomly adjusting the pitch value within a range of -3 to 3 semitones, where positive values raise the pitch and negative values lower it[9,10]. Noise injection is applied by adding a uniform magnitude of 0.005 across all time frames, effectively introducing white noise[7]. Time stretching adjusts the playback speed at a randomized rate between 0.8 and 1.2, with values greater than 1.0 accelerating the audio and values below 1.0 slowing it down[9,10]. Time shifting introduces a random temporal adjustment within a range of -0.5 to +0.5 seconds, which is then converted into samples and applied accordingly. The echo effect is implemented by applying a delay of 0.25 seconds with an attenuation factor randomly selected between 0.2 and 0.3. The audio is processed using a filter to simulate the echo and the resulting signal is normalized to maintain audio quality. The reverb technique involves generating an impulse response with a duration between 0.1 and 0.3 seconds, which is then applied to the audio via convolution. A reverb strength of 0.4 is used and the audio is normalized to avoid distortion or clipping. Finally, background noise is introduced by mixing the original audio with background sounds randomly selected from a predefined library of noise samples. The background audio is repeated to match the length of the original sample and a volume factor of 0.5 is applied to ensure the noise does not overpower the original speech.

B. Spectrogram-based Augmentations

We apply data augmentation using three variations of masking in SpecAugment[11] (SA), considering time, frequency and time-frequency masking, as well as MixSpeech (MS)[12].

SA in the frequency domain features[11] focuses solely on frequency masking, omitting the time domain to isolate the effects of frequency-based augmentation. The log-mel spectrogram is divided into segments along the vertical axis (frequency/mel-bins) and masks are applied randomly. In this study, six masks are applied to the mel-bins, with mask dimensions selected randomly. Each mask coveres between 6% and 9% of the total mel-bins, effectively obscuring portions of the frequency spectrum. By excluding time domain masking, this approach highlights the unique effects of frequency-based transformations on model performance. To further augment the training data, SA in the time domain features is also employed[11]. This technique involves masking random sections of the audio along the time axis of the log-mel spectrogram, effectively creating gaps in the temporal dimension. In this study, 20times masks are applied across the log-mel spectrogram, with mask sizes ranging from 2% to 3% of the total time frames. This approach helps the model learn to be invariant to missing or corrupted sections of the input audio, enhancing robustness and generalization capability. By focusing on the time domain, this augmentation avoids altering the frequency content, ensuring the integrity of the audio's spectral characteristics. Finally we apply the standard SA technique, which involves masking random sections of the log-mel spectrogram across both time frames and mel-bins. To adapt the original SA method[11], the log-mel spectrogram is divided into distinct segments along the horizontal (time) and vertical (frequency) axes. In this study, 20 masks are placed across the time frames and six masks are applied to the mel-bins. The dimensions of the masks are randomly selected, with the time masks covering 2% to 3% of the total time frames and the mel-bin masks covering 6% to 9% of the mel-bins.

MS12 involves combining spectrograms from different speech samples along with their corresponding labels. In this method, two samples are merged by overlaying the spectrogram of one sample onto another while preserving the original labels. Due to the complexity of incorporating label mixing into the sequence-to-sequence ASR model, label mixing is avoided before training to prevent ambiguity in the model. The log-mel spectrograms of both samples are obtained using the Whisper Feature Extractor5 (WFE). The spectrograms are combined by applying an opacity value of 0.8 to the base sample and 0.2 to the random sample, ensuring precision in the merged spectrograms.

C. Combining Augmentation Techniques: A Sequential Strategy

In this approach, combined augmentations are applied to enhance the diversity of the training data by utilizing multiple methods sequentioaly. The combinations are structured as follows: AbA with SA, AbA with MS and MS with SA. Each combination uses consistent hyperparameters to ensure uniformity throughout the augmentation process.

In AbA+SA combination, AbAs are applied first to the raw audio samples. Afterward, SA (using frequency and time masking) is applied to the generated log-mel spectrograms to introduce additional variability. By applying the audio-based transformations first, we ensure that any variations in the raw audio are propagated through the spectrogram, allowing SA to operate on a richer set of features. For AbA+MS combination, AbAs are again applied first to modify the raw audio samples. Following this, MS is applied, mixing multiple samples together to simulate overlapping speech scenarios. This combination helps create more challenging training conditions for the ASR model, improving its ability to handle real-world audio situations. Finally in the MS+SA combination, MS is applied first, combining multiple samples to create overlapping speech. Afterward, SA (with both frequency and time masking) is applied to the generated mixed log-mel spectrogram.

3. Experimental Setup

A. Dataset

The Common Voice 17 dataset[13] is selected for this study due to its open-source availability, multilingual composition and diverse characteristics. This dataset contains audio samples paired with corresponding transcripts, with durations ranging from 1 to 15 seconds. The data quality varies significantly, featuring both clear, high-quality recordings and samples with substantial background noise, making it an ideal choice for training robust ASR models across diverse conditions.

For this study, four languages are utilized: German (DE), Russian (RU), Farsi (FA) and Arabic (AR). A filtering process is applied using the Whisper tokenizer to exclude any samples from the original dataset where the token indices sequence length exceeded the model's maximum limit of 1024, as defined for the Whisper small model. After filtering, a subset of 100,000 paired audio-transcript samples for each language is selected for training. Additionally, 10% of the training sample size for each language is allocated for the validation set and another 10% for the test set. The data splitting approach followed the guidelines outlined in Joseph et al.[14]. The entire test set is subsequently used to assess the model's performance during the final evaluation.

B. Adjustment of Data Augmentation Methods

The adjustment process begins with the pre-processing stage, which involves standardizing the audio samples to a consistent sampling rate of 16,000 Hz, the required input format for the Whisper model. After resampling, non-essential labels are removed, leaving only the necessary transcription data. Next, WFE5 is employed to compute log-mel spectrogram features for each audio sample. Since the Whisper model expects input segments of 30 seconds, all audio samples are either padded or truncated to this length to maintain consistency. Similarly, padding is applied to the transcription labels to ensure they match the padded audio samples.

Data augmentation methods are then applied to enhance the diversity of the training dataset. AbA techniques are randomly selected for each sample, with one of seven available methods chosen to avoid overwhelming the model with simultaneous augmentations that could hinder learning. These augmentations are applied directly to the audio samples before feeding them into the WFE. After computing the log-mel spectrograms using WFE module, spectrogram-based augmentations are applied. The augmented data is subsequently used for the training split, while the validation set remains unaltered to ensure a consistent evaluation of the model's performance.

C. Training Procedure

OpenAI's Whisper models are available in various sizes, ranging from tiny to large. For this investigation, Whisper model in the small multilingual configuration is used[5]. This model is selected due to its moderate capacity, providing a balance between computational efficiency and performance, as well as its proven effectiveness in prior research5. Its ability to handle multiple languages made it particularly suitable for expanding research beyond the English-centric scope typical in many ASR studies[15]. The training process begin by establishing the baseline model, trained on the labeled dataset without any augmentation. This baseline served as a reference point to evaluate the impact of different augmentation techniques on the model performance, identifying which strategies provided the most significant improvement.

The training configuration is optimized as follows: All models are trained with weight decay of 0.1 and learning rate of 1e-5 using standard Adam optimizer[16] for up to 50,000 steps, which is approximately equivalent to 10 epochs. Additionally, gradient accumulation is implemented to simulate larger batch training, with an effective batch size of[2]. This setup provides a structured framework for evaluating the impact of different augmentation strategies on the performance of the Whisper small multilingual model.

D. ASR Performance Evaluation Across Languages

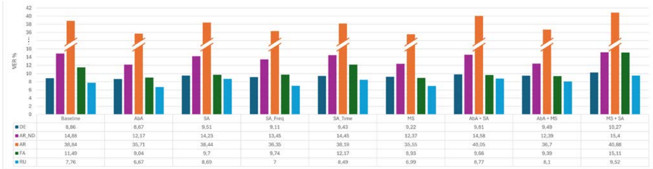
The Word Error Rate (WER)[17] is the standard metric used to evaluate the performance of ASR models across low-resource languages, including German (DE), Arabic without diacritics (AR_ND), Arabic with diacritics (AR), Farsi (FA) and Russian (RU). This metric measures transcription accuracy by calculating the number of substitutions, deletions and insertions required to match the model's output to a reference transcript. This provides a comprehensive assessment of transcription quality by capturing various types of errors.

For Arabic, two separate evaluations are performed: one with diacritics preserved and another with diacritics removed from both transcripts and predictions. This distinction is important because some transcripts in the dataset lack diacritics, which could unfairly penalize models that predict with diacritics. By removing diacritics in one test set, we ensured a more balanced and fair evaluation of model performance.

4. Results

Figure 1 Presents our experimentation results on the test set. In this figure, the performance comparison of different augmentation strategies compared to the baseline reveals several important trends in the WER[17] across low-resource languages.

The AbA demonstrates exceptional performance, leading to the improvements across all languages. Notably, it is the top-performing method, achieving the most significant WER reductions compared to competing augmentation approaches, with relative improvements of 2.14%/18.21%/14.05% over the baseline in DE/AR_N/RU. Additionally, AbA shows relative improvements of 8.06% and 21.32% for AR and FA, respectively. Among spectrogram-based approaches, MS outperforms SA, showing improvements over the baseline for all languages except DE. However, it generally lags behind AbA in AR_ND and RU. FA and AR, on the other hand, benefit significantly from MS compared to the other languages, with relative improvements of 22.28% and 8.47%, respectively. This indicates that MS is particularly effective for FA, while its benefits for other languages are more modest. The SA technique is evaluated using three different masking approaches, as explained in Section 3. The best performance comes from SA utilized freqauncy masking (SA_Freq), which, although outperforming the baseline for all languages except DE, its impact is less pronounced than AbA and MS augmentations.



	Baseline	AbA	SA	SA_Freq	SA_Time	MS	AbA + SA	AbA + MS	MS + SA
■ DE	8,86	8,67	9,51	9,11	9,43	9,22	9,81	9,49	10,27
■ AR_ND	14,88	12,17	14,23	13,45	14,45	12,37	14,58	12,39	15,4
■ AR	38,84	35,71	38,44	36,35	38,19	35,55	40,05	36,7	40,88
■ FA	11,49	9,04	9,7	9,74	12,17	8,93	9,66	9,39	15,11
■ RU	7,76	6,67	8,69	7	8,49	6,99	8,77	8,1	9,52

Figure1: Performance comparison of the baseline and models using different augmentation techniques in terms of WER% across four languages on the test set.

Figure 1 further examines the performance when combining augmentation methods as described in Section 2. The results from these combined strategies are substantial. Combining Top-2 augmentation techniques, i.e. AbA with MS shows the best performance within the combined augmentations, with relative improvement of 16.73%/5.51%/18.28% over the baseline for AR_N/AR/FA languages. However, the gains are less substantial than when AbA or MS are used alone, suggesting diminishing returns when both methods are applied simultaneously. AbA+SA combination yields moderate improvements in WER. While the benefits are clear for AR_N/FA languages with relative reduction of 2.02%/ 15.93% over the baseline, the results again indicate diminishing returns from the simultaneous application of both methods. This suggests that the masking effect of SA complements the noise introduction from AbA, but the combined impact may not be universally beneficial for all languages. For DE and RU, where SA alone failed to improve performance over the baseline, combining it with AbA does not yield any further gains. Finally MS+SA shows the worst performance among all standalone and combined augmentations, with no improvements in terms of WER for any language.

5. Augmentation Effects and Model Behavior

This paper presents a comparative analysis of various data augmentation methods used for ASR, evaluating their performance and generalization ability. In this section, we address the research questions stated in the introduction by assessing model behavior and performance on the validation set, as shown in **(Figure 2 and Figure 3)**.

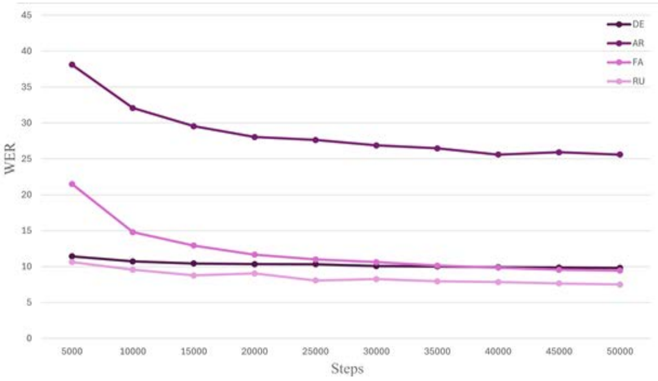


Figure 2: Learning curves of various baseline models across languages, measured on the validation set in terms of WER%.

The baseline performance in **Figure 2**, which serves as a reference point, indicates that the model performs better on FA and RU, with lower WER values compared to AR and DE, where the WER is particularly high for AR at 38.11%, suggesting the model initially struggles more with this language. In addition to the baseline, **Figure 3** illustrates the impact of three different augmentation methods on both standalone **(c.f. Figure 3(a))** and combined **(c.f. Figure 3(b))** strategies. This comparison provides insights into the relative effectiveness and generalization ability of each technique. The general trend which clearly is obvious from the experiments is consistent improvement of all models throughout the training period, with a significant boost in performance around the regulated 10-epoch mark. Notably, the model fine-tuned using AbA achieved the highest validation performance across DE/AR/RU languages as depicted in **Figure 3(a)**. The strength of AbA lies in its ability to directly manipulate raw audio samples, ensuring that even subtle and nuanced sounds are accurately captured. This augmentation technique enhances the model's robustness by improving its ability to generalize across varied and ambiguous audio data.

A key feature of AbA is its capacity to simulate real-world conditions, incorporating environmental noises such as traffic sounds. This helps the model to learn how to handle speech segments with low clarity or noise interference, similar to masking effects, where certain parts of speech are obscured. The augmentation can be represented by a transformation scale ranging from -1 to 1, where 0 represents clear speech, 1 represents unintelligible speech and values closer to -1 represent noise-dominated samples. Audio augmentation effectively shifts this spectrum towards intelligible speech, helping the model to perform better under noisy conditions. In conclusion, the results indicate that AbA proves to be the most effective technique overall with consistently outperforms the baseline across all languages.

A. Regularization Affect

Data augmentation strategies play a vital role not only in improving model performance but also in preventing overfitting, acting as a form of regularization. Regularization refers to techniques that prevent models from memorizing training data, thereby enhancing their ability to generalize to unseen data. Different augmentation techniques exhibit varying degrees of regularization, which directly influences the WER across training steps.

B. Overfitting Indicators

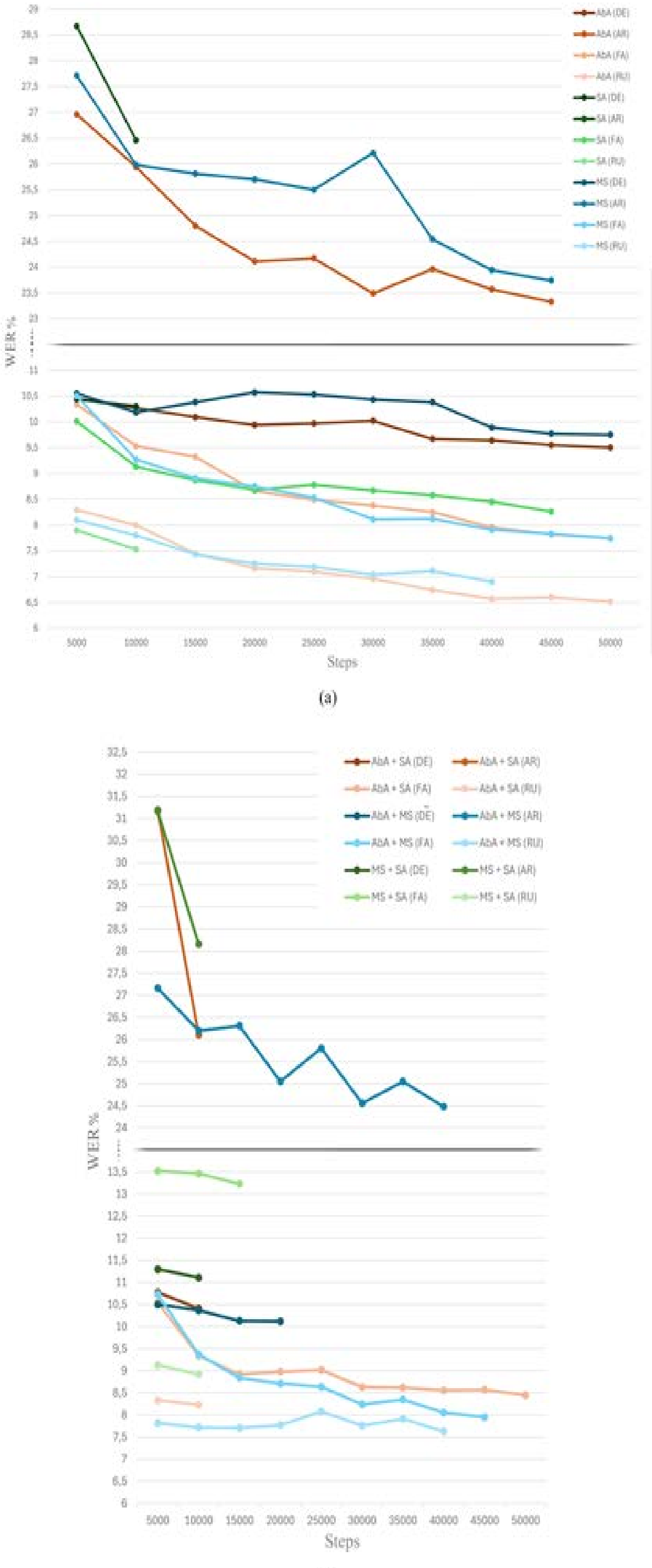
With data augmentation, it is also possible to avoid overfitting by injecting various alterations into the training data. This prevents the model from learning the same data excessively and reduces the risk of overfitting[18]. Among the augmentation techniques tested, AbA demonstrated the greatest efficacy in preventing overfitting. As shown in the **Figure 3 (a)**, this method consistently reduced WER across multiple languages and training steps. Even after extended training periods, models using AbA maintained stable and lower WER values compared to other methods, such as SA and MS. This suggests that AbA introduces a balanced variety of perturbations in the audio samples, preventing the model from overfitting to the specific training set.

AbA directly manipulates audio signals, which contributes to a more robust learning process by forcing the model to adapt to variations that mimic real-world conditions. This continuous introduction of variability prevents the model from over-relying on specific patterns, enhancing its ability to generalize to new and unseen audio inputs. The improvements in WER observed for languages like AR and FA further emphasize this point, with significant reductions in WER when compared to the baseline.

C. Stability in Learning

Incorporating an effective regularization technique can significantly improve the overall effectiveness of the learning process**17**. The stability of AbA is reflected in its ability to produce consistent WER reductions throughout training. Unlike SA or MS, which demonstrated some variability across languages and epochs, as depicted in **Figure 3(a)**, AbA maintained a steady performance across both lower and higher training steps. This stability indicates that the augmentation technique not only provides effective regularization but also allows the model to learn meaningful features from the data without overfitting.

Moreover, the combination of AbA particularly with MS **(c.f. Figure 3(b))**, continue to show improvements in generalization, though the individual performance of AbA stay out as the most effective. The ability of this technique to consistently reduce WER across a diverse set of languages demonstrates its superior regularization effect, particularly for challenging languages like AR.



(a)

(b)

Figure3: Learning curves of different augmentation methods across languages, measured on the validation set in terms of WER%. (a) Standalone augmentations, (b) Combined augmentation strategies.

D. Integrating Insights from Cross-Linguistic Data Augmentation for Deep Learning Models

Insights from data augmentation techniques, particularly in multilingual contexts, highlight the need for tailored strategies that accommodate cross-linguistic variations. Languages differ in phonetic structures, acoustic properties and script complexities, which can impact how augmentation techniques influence model performance. Understanding these variations is key to designing augmentation strategies that generalize well across languages, as different methods yield different levels of success depending on the target language.

AbA indicates the most universally effective, particularly in languages with more complex phonological and acoustic variations, such as AR and FA. This method improved model performance by introducing variability in the acoustic features, simulating real-world conditions and enhancing robustness across diverse linguistic environments (see **Figure 3**). The improvement is especially pronounced for AR, where the WER dropped significantly, highlighting the role of augmentation in generalization. As shown in **Figure 3(a)**, among spectogram-based augmentation techniques, MS demonstrates greater generalization ability and better performance on validation set compared to SA. This method, which blends multiple audio streams, proves particularly effective for FA, where the WER continued to decline even in later epochs, achieving the relative improvement of 18.01% over the baseline. Across the other -AR, DE and RU-MS also enhances the model's generalization ability, with relative reductions in WER of 7.23%, 0.61% and 8%, respectively.

From these findings, it is clear that data augmentation techniques must be carefully adapted based on linguistic characteristics. For instance, AbA provides consistent improvements across languages, while SA and MS show varying results depending on the phonetic and acoustic properties of the target language. This suggests that cross-linguistic insights into data augmentation can inform more sophisticated deep learning models by adjusting augmentation strategies to suit specific language families. Furthermore, our findings reveal different behavior of these techniques not only on the learning process but also on model regularization and generalization. This insight is crucial for integrating these methods into more sophisticated deep learning-based augmentation approaches, particularly those using aligned data augmentation in conjunction with classic approaches, such as SA in On-the-Fly[19] or where an AbA pipeline is used in parallel with text augmentation in TGSS[20] method.

6. Conclusion

This study undertook a comprehensive evaluation of various data augmentation techniques to improve the Whisper model's performance for low-resource languages. The investigation revealed that augmentation methods, including time, frequency and time-frequency domain features, significantly impact the model's generalization ability and representation improvement across different languages.

The findings indicate that AbA techniques, which manipulate audio samples directly, offer substantial benefits in enhancing the model's adaptability to variable audio inputs. These techniques demonstrated notable effectiveness in improving model performance, particularly in languages with diverse acoustic characteristics. Such improvements align with previous research that emphasizes the efficacy of AbA in similar contexts[4,19]. Furthermore, the study highlighted that the effects of data augmentation techniques on learning behavior, model regularization and generalization vary across languages. This variation underscores the importance of tailoring augmentation strategies to specific linguistic contexts to optimize model performance. These insights are consistent with findings from other studies on the impact of augmentation on model behavior and performance[20].

Insights from this research suggest that integrating advanced deep learning-based augmentation methods with traditional approaches could further enhance model efficiency and effectiveness. Such integration promises to generate diverse training samples, improve the cost-effectiveness of training larger models and offer greater flexibility in applying various augmentation techniques. Overall, these findings contribute valuable knowledge on leveraging data augmentation to address the challenges of low-resource language processing, paving the way for future advancements in ASR models.

7. References

1. Kshirsagar S, Pendyala A, Falk TH. Task-specific speech enhancement and data augmen-tation for improved multimodal emotion recognition under noisy conditions. Frontiers in Computer Science 2023;5.
2. Coto-Solano R, Nicholas SA, Datta S, et al. Development of Automatic Speech Recog- nition for the Documentation of Cook Islands M¯aori. in Proceedings of the Thirteenth Language Resources and Evaluation Conference (Marseille, France):European Language Resources Association 2022;3872-3882.
3. Guillaume S, Wisniewski G, Macaire C, et al. Fine-tuning pre-trained models for Au- tomatic Speech Recognition, experiments on a fieldwork corpus of Japhug (Trans-Himalayan family). in Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages(Dublin, Ireland): Association for Computational Linguistics 2022;170-178.
4. Bartelds M, San N, McDonnell B, Jurafsky D, Wieling M. Making More of Little Data: Improving Low-Resource Automatic Speech Recognition Using Data Augmentation 2023.
5. Radford A, Kim JW, Xu T, Brockman G, McLeavey C, Sutskever I. Robust Speech Recognition via Large-Scale Weak Supervision 2022.
6. Ko T, Peddinti V, Povey D, Seltzer ML, Khudanpur S. A Study on Data Augmenta- tion of Reverberant Speech for Robust Speech Recognition. in 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP):5220-5224IEEE Press 2017.
7. Kathania HK, Kadiri SR, Alku P, Kurimo M. Using Data Augmentation and Time- Scale Modification to Improve ASR of Children's Speech in Noisy Environments. Applied Sciences 2021;11(18):8420.
8. Wei S, Zou S, Liao F, Lang A Comparison on Data Augmentation Methods Based on Deep Learning for Audio Classification. Journal of Physics: Conference Series 2020;1453(1):012085.
9. Wei S, Zou S, Liao F, Lang W. A Comparison on Data Augmentation Methods Based on Deep Learning for Audio Classification. Journal of Physics: Conference Series 2020;1453:012085.
10. McFee B, Raffel C, Liang D, et al. librosa: Audio and Music Signal Analysis in Python. in librosa: Audio and Music Signal Analysis in Python 2015;18-24.
11. Park DS, Chan W, Zhang Y, et al. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. in Interspeech 2019ISCA 2019;1.
12. Meng L, Xu J, Tan X, Wang J, Qin T, Xu B. MixSpeech: Data Augmentation for Low-resource Automatic Speech Recognition 2021.

13. Ardila R, Branson M, Davis K, et al. Common Voice: A Massively-Multilingual Speech Corpus 2020;1.

14. Joseph VR, Vakayil A. SPlit: An Optimal Method for Data Splitting. Technometrics 2021;64(2):166-176.

15. Makhijani R, Shrawankar U, Thakare VM. Opportunities Challenges In Automatic Speech Recognition. 2013.

16. Kingma, D. P., Ba, J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2017.

17. Morales N, Hansen J, Toledano D. MFCC Compensation for Improved Recognition of Filtered and Band-Limited Speech. in MFCC Compensation for Improved Recognition of Filtered and Band-Limited Speech1 2005;521- 524.

18. Ying X. An Overview of Overfitting and its Solutions. Journal of Physics: Conference Series. 2019;1168:022022.

19. Lam TK, Ohta M, Schamoni S, Riezler S. On-the-Fly Aligned Data Augmentation for Sequence-to-Sequence ASR. in Interspeech 2021ISCA 2021;1.

20. Huang Z, Keren G, Jiang Z, et al. Text generation with speech synthesis for ASR data augmentation. arXiv. 2023.