International Research Journal on Advanced Engineering Hub
e ISSN: 2584-2137
Vol. 04 Issue: 01 January 2026
Page No: 111 - 117
www.irjaeh.com
https://doi.org/10.47392/IRJAEH.2026.0014

Wav2Vec-based Audio Data Augmentation for Low-Resource Speech Recognition

P. Haritha[1]*, P. Shanmugavadivu*[2]
[1,2]*Department of Computer Science and Applications, The Gandhigram Rural Institute (Deemed to be University), Gandhigram,Tamil Nadu, India.*
Emails: *7haricsa@gmail.com*[1]*, psvadivu@ruraluniv.ac.in*[2]

Abstract

Audio Data Augmentation (ADA) is a transformative process of small datasets into voluminous datasets. ADA can be performed on any type of dataset namely Images (Mel Spectogram), Audio and Text, based on applications such as Gender Identification, Speech Recognition, Text Summarization. ADA plays a vital role in the development of Automatic Speech Recognition (ASR) systems when the experimental language datasets are smaller in size and are being low resource. This article focuses on performing ADA techniques namely the addition of noise, pitch shifting, increasing or decreasing of speed and adding reverberation to the audio signals. The proposed method includes preprocessing, data augmentation, audio transcription using pre-trained Self-Supervised Learning based Wav2vec models; and finally with the post-processing of data on the removal of induced tags from the transcribed data. The article integrates audio transcription after performing audio augmentation techniques to evaluate the quality of speech using Word Error Rate (WER). The proposed Audio Data Augmentation for Low Resource Speech Recognition (ADA-LRSR) with the integration of Wav2Vec (Vakyansh) achieved an overall WER of 0.5231, which was promising than that of other Wav2Vec variants (Base and Large). The suggested approach is evaluated on a manually recorded 39 preprocessed audio files and obtained 312 audio files after augmentation. In addition, ADA-LRSR's framework chose the addition of noise and reverberation as the best augmentation techniques with preservation of speech quality.

Keywords: *Audio Data Augmentation, Low-Resource Speech Recognition, Wav2Vec, Automatic Speech Recognition, Audio Processing.*

1. Introduction

In recent years, the utilization of Automatic Speech Recognition (ASR) for languages in huge volume with transcribed speeches are being made available. In addition, Self-Supervised Learning (SSL) based pre-trained models are trained with huge volume of data such as 960 hours of speech. In general, the SSL models require a reasonably large dataset for effective learning of latent patterns. In order to mitigate this issue of learning on low resource languages, they are combined with the original datasets to mitigate the data scarcity [1,2] This article focuses on increasing the volume of input dataset with a due concern for the preservation of quality of speech after augmentation. The proposed work is designed to perform augmentation using proven methods along with a post-processing component to perform transcription, resulting in the best suitable augmentation techniques for speech processing applications.

2. Related Works

The researchers [3] achieved an improved performance of ASR systems after augmenting 24 minutes of manually transcribed data from four Germanic Language variants Gronings, West-Frisian, Besemah and Nasal that are considered as low resource. It is evident that after augmentation the WER is decreased from 53.3% to 30.1%. The article [4] utilized two levels of augmentation, namely signal-based augmentation and speaker-based augmentation on four languages Amharic, Guarani, Igbo and Pashto from Intelligence Advanced Research Projects (IARPA). The signal-based augmentation deals with Noise, Speed Perturbation and Reverberation. The speaker-based augmentation utilized feature Space Maximum Likelihood Linear Regression (fMLLR) transform to incorporate other speakers' speech characteristics to the original speaker's audio. Out of the two levels, signal-based augmentation provided reduced WER by 1.2% whereas speaker-based augmentation reduced WER by 0.6% and also the technique requires additional computational costs. In article [5], the speed perturbation technique with the factors of 0.9, 1.0 and 1.1 to augment the audio signals is used in order to avoid overfitting as well as to improve the efficiency of Deep Neural Network (DNN) models. The techniques were evaluated with 100 hours to 960 hours of speech data from GALE Mandarin, Tedlium, Switchboard, Librispeech and ASpIRE resulted in an average relative improvement of 4.3% after implementing the speech perturbation. The authors [6] have augmented single speaker audio files using Pitch-Speed Feature Space (PSFS) with its variations and evaluated the results with an improvement of 10.3% in ASR's accuracy. The researchers [7] investigated the impact of noise addition and modifications in speech. It was confirmed that data augmentation has contributed to improve ASR on Latin American and Asian accented English speech, as measured in terms of WER reductions up to 30%.

3. Motivation

The increased use of Self-Supervised Learning models with their consistent performance has motivated the researchers to explore the possibility of training them in low resource datasets, while designing automated speech recognition systems. This research is designed to apply ADA on low resource audio datasets and then check their usability on the pre-trained models of ASR.

Audio Data Augmentation for Low-Resource Speech Recognition (ADA-LRSR) (Proposed) The proposed method performs ADA on manually recorded Indian Accent Audio files to integrate the pre-trained ASR model Wav2Vec and its versions. ADA-LRSR incorporates two phases namely Data Augmentation and ASR-based Transcription to select the best augmentation techniques that preserve the quality of speech. In Fig.1 the flow diagram of the proposed ADA-LRSR is given.

3.1. Data Augmentation

ADA-LRSR uses four prominent augmentation techniques namely Time Stretching, Pitch Shifting, Addition of Gaussian White Noise and Reverberation on manually recorded Indian accented audio files that are resampled from 44.1 kHz to 16 kHz. The process is referred to be deformations of audio files. The resampled audio files are fed into ASR models. The audio files are also converted into mp3 to wav format which is meant to be the raw audio format used for processing the audio data. Thus, each audio files are made up of 5 sets of augmented audio data along with the original audio files. The description of the chosen four augmentation methods for the audio data are given herein under.

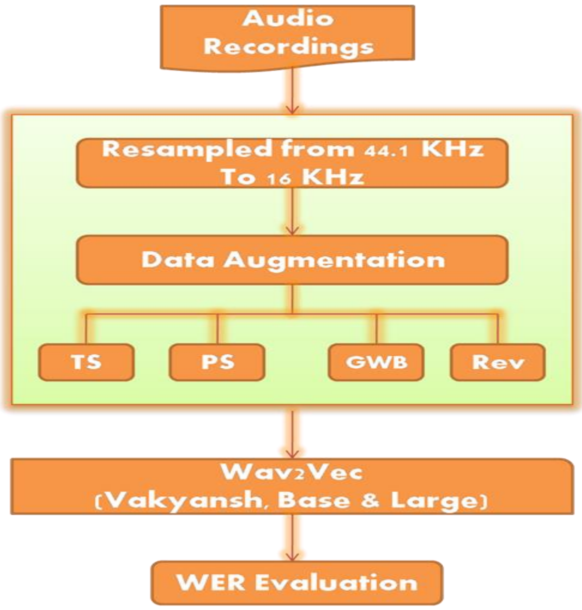
Audio Recordings
Resampled from 44.1 KHz To 16 KHz
Data Augmentation
TS
PS
GWB
Rev
Wav2Vec (Vakyansh, Base & Large)
WER Evaluation

Figure 1 **Flow Diagram of ADA-LRSR**

Time Stretching (TS): The technique can either accelerate or decelerate the audio signals referred to as Speech Perturbation. TS uses the time stretching factors {0.9, 1.1}, resulting in speeding or slowing down the speeches present in the audio files preserving the pitch. TS adjusted only 10% of the stretch to preserve the quality of speech [8, 9].

Pitch Shift (PS): The technique focuses on adjusting the pitches of the audio signals, preserving the duration of audio signals. PS adjusts the semitone as {+2 and -2} to reproduce the audio signals with higher and lower pitches [10,11,12].

Gaussian White Noise (GWN): The technique adds Gaussian white noise to the audio signals with a factor of 0.005 that simulates background noise of low intensity that suits ASR with the assurance for preservation of the quality of speech [13].

Reverberation (Rev): The technique feeds multiple reflections in the form of echoes. Rev feeds short and long reverb with the factors {0.2 and 0.4} that induce the effect of closed environment such as small room, temple, hall etc. This technique is often referred to be Room Impulse Response (RIR) [14, 15].

3.2. ASR based Transcription

The phase aims to transcribe the augmented audio files using self-supervised pre-trained models of ASR's Wav2Vec and versions (Base, Large and Vakyansh) that are pre-trained with a huge volume of unsupervised data that can be fine-tuned with small-sized supervised data [16, 17]. The pre-trained Wav2Vec model gets trained on the augmented audio samples in WAV format which are resampled in the 16 kHz along with the CSV file that consists of the transcripts as ground truth. The versions of Wav2Vec were pre-trained on 960 hours of speech audio files with transcriptions. Initially, the conversion of stereo to mono channel is performed. The Wav2Vec processor extracts the features from raw audio wave forms and converts them into tensors that are fed to Wav2VecforCTC, which performs speech recognition using Connectionist Temporal Classification (CTC) and provides them as logits which are unnormalized prediction scores for transcriptions. Argmax is fed with logits to decode the predicted tokens and obtain the final transcription using CTC decoder. The post- processing covers the removal of induced tag from the transcribed data introduced by the pre-trained models, conversions of lower-case letters and removal of special symbols preserving the spaces between the words. The final transcribed data is stored in CSV file format. The speech quality of transcribed audio data was evaluated.

3.2.1. Dataset Description & Evaluation

ADA-LSR uses manually recorded 39 audio files of Indian Accent for the evaluation. The dataset includes a mono speaker's voice recorded in a controlled environment through the smartphone Redmi Note 7. The dataset was converted from mp3 to wav format and also the dataset was preprocessed for Silence removal using Optimized Voice Activity Detection [18,19]. CSV file was created to document the transcripts of the recorded 39 audio files. The transcripts are considered as ground truth for further performance evaluation. The quality of speech was, evaluated using Eqn (1),

$$\boldsymbol{WER} = \frac{\boldsymbol{S+D+I}}{\boldsymbol{N}} \qquad \textbf{(1)}$$

where S denotes substitutions, D denotes deletions, I for deletion and N denotes the number of words in reference. The computation of WER on transcribed text obtained from augmented audio files was done using Python's library jiwer by keeping the ground truth transcripts as reference. The lower the WER, the better the speech quality.

3.2.2. Results and Discussions

The manually recorded 39 files were augmented to 312 audio files out of which those transcriptions with a threshold of WER less than or equal to 0.8 were considered as the best augmentation techniques. The transcriptions were categorized to Very Good, Good, Moderate and Acceptable based on WER range of [0.0, 0.2], [0.2, 0.4], [0.4, 0.6] and [0.6, 0.8] respectively. The following Table 1 & 2 determines that Vakyansh based Wav2Vec model that is pre-trained with Indian Accented speech is performing well than the other pre-trained models chosen for evaluation.

Table 1 **Average WER on Data Augmentation Techniques**

Pre-Trained Models	Augmentation Techniques	Average WER
Vakyansh-Wav2Vec	Noise	0.3571
	PS (-2)	0.6000
	PS (+2)	0.6778
	Rev (0.2)	0.5444
	Rev (0.4)	0.5000
	TS (0.9)	0.6500
	TS(1.1)	0.6556
	Overall WER	0.5231
Wav2Vec Base	Noise	0.5000
	PS (-2)	0.7000
	PS (+2)	0.7000
	Rev (0.2)	0.7000
	Rev (0.4)	0.7000
	TS (0.9)	0.7000
	TS (1.1)	0.7000
	Overall WER	0.5957
Wav2Vec Large	Noise	0.4467
	PS (-2)	0.6333
	PS (+2)	0.6000
	Rev (0.2)	0.6556
	Rev (0.4)	0.6400
	TS (0.9)	0.7000
	TS (1.1)	0.5000
	Overall WER	0.5683

Table 2 **Categorization of Audio Files based on WER**

Pre-Trained Models	Augmentation Techniques	0.0 – 0.2 (Very Good)	0.2-0.4 (Good)	0.4-0.6 (Moderate)	0.6-0.8 (Acceptable)
Vakyansh-Wav2Vec	Noise	8	6	12	2
	PS (-2)	0	1	2	5
	PS (+2)	0	0	1	8
	Rev (0.2)	0	2	10	6
	Rev (0.4)	2	2	10	6
	TS (0.9)	0	0	3	9
	TS (1.1)	0	02	2	7
	Total (105)	10	12	40	43
Wav2Vec Base	Noise	2	2	2	6
	PS (-2)	0	0	0	1
	PS (+2)	0	0	0	1
	Rev (0.2)	0	0	0	3

	Rev (0.4)	0	0	0	4
	TS (0.9)	0	0	0	1
	TS (1.1)	0	0	0	1
	Total (23)	2	2	2	17
Wav2Vec Large	Noise	3	2	6	4
	PS (-2)	0	0	1	2
	PS (+2)	0	0	1	1
	Rev (0.2)	0	0	2	7
	Rev (0.4)	0	1	1	8
	TS (0.9)	0	0	0	1
	TS (1.1)	0	0	1	0
	Total (41)	3	3	12	23

The above table shows that the augmentations of audio files are considerable only if they preserve the quality of speech. From the above given tables it is also notable that in addition with augmentation techniques pre-trained model plays a vital role in transcribing texts. Table 1 resulted in 0.5231 as the overall WER of Wav2Vec's Vakyansh performed better than the other variants Base and Large WER: 0.5957 and 0.5683. Table 2 shows that 105 audio files out of 312 augmented files provided WER as 0.0 to 0.2 for 10 audio files, 0.2 to 0.4 for 12 audio files, 40 audio files of 0.4 to 0.6 in WER and WER of 0.6 to 0.8 from 43 audio files. In addition to the augmentation techniques such as noise and Rev, quality speech is provided after augmentation.

Conclusion

The article aims to perform audio data augmentation on Indian accented voice datasets and also addresses the data scarcity issue. The proposed approach combines data augmentation and ASR-based audio transcriptions to choose better augmentation techniques and best suited version of Wav2Vec that depends on Self-Supervised Learning. The article ensures the quality of speech even after the augmentation of audio signals using noise addition, manipulation of pitch & speed and inducing reverberation. The article concluded that noise and reverberation along with the Vakyansh based Wav2Vec model are providing better results than other augmentation techniques as well as the versions of Wav2Vec with an overall WER of 0.5231. The article also highly recommends that the augmented transcripts with WER from 0.0 to 0.6 also provide good results. Therefore, such augmented transcripts can also use for further Speech Recognition-based tasks.

Future Enhancement

The article has narrowed down to focus on the augmentation techniques with respect to speech recognition process. It led a path to the adjustments of factors involved in augmentation techniques to analyze the quality of speech. Thus, the article leads to further improvements in the augmentation techniques along with the pre-trained models to multiply the datasets into several numbers that will be helpful for Low-Resource Speech Languages.

Acknowledgement

The authors gratefully acknowledge the financial support by National Fellowship for Other Backward Classes (NFOBC) scheme of Ministry of Social Justice & Empowerment and computational facilities granted by UGC-Special Assistance Programme (SAP), Image Processing Laboratory, Department of Computer Science and Applications, The Gandhigram Rural Institute (Deemed to be University).

References

[1]. Goodfellow, Ian J., Mehdi Mirza, Da Xiao, Aaron Courville, and Yoshua Bengio. "An empirical investigation of catastrophic

forgetting in gradient-based neural networks." arXiv preprint arXiv:1312.6211 (2013).

[2]. Coto-Solano, R., Nicholas, S. A., Datta, S., Quint, V., Wills, P., Powell, E. N., ... & Feldman, I. (2022, June). Development of automatic speech recognition for the documentation of Cook Islands Māori. In Proceedings of the Thirteenth Language Resources and Evaluation Conference (pp. 3872-3882).

[3]. Bartelds, M., San, N., McDonnell, B., Jurafsky, D., & Wieling, M. (2023). Making more of little data: Improving low-resource automatic speech recognition using data augmentation. arXiv preprint arXiv:2305.10951.

[4]. Hartmann, W., Ng, T., Hsiao, R., Tsakalidis, S., & Schwartz, R. M. (2016, September). Two-Stage Data Augmentation for Low-Resourced Speech Recognition. In Interspeech (pp. 2378-2382).

[5]. Ko, T., Peddinti, V., Povey, D., & Khudanpur, S. (2015, September). Audio augmentation for speech recognition. In Interspeech (Vol. 2015, p. 3586).

[6]. Zahid, S. M., & Qazi, S. A. (2025). Pitch-Speed Feature Space Data Augmentation for Automatic Speech Recognition improvement in Low-Resource Scenario. IEEE Access.

[7]. Fukuda, T., Fernandez, R., Rosenberg, A., Thomas, S., Ramabhadran, B., Sorin, A., & Kurata, G. (2018, September). Data Augmentation Improves Recognition of Foreign Accented Speech. In Interspeech (No. September, pp. 2409-2413).

[8]. Salamon, J., & Bello, J. P. (2017). Deep convolutional neural networks and data augmentation for environmental sound classification. IEEE Signal processing letters, 24(3), 279-283.

[9]. Yuanchao, X., Zhiming, C., & Xiaopeng, K. (2023). Improved pitch shifting data augmentation for ship-radiated noise classification. Applied acoustics, 211, 109468.

[10]. 10. Schlüter, J., & Grill, T. (2015, October). Exploring data augmentation for improved singing voice detection with neural networks. In ISMIR (pp. 121-126).

[11]. Terashima, R., Yamamoto, R., Song, E., Shirahata, Y., Yoon, H. W., Kim, J. M., & Tachibana, K. (2022). Cross-speaker emotion transfer for low-resource text-to-speech using non-parallel voice conversion with pitch-shift data augmentation. arXiv preprint arXiv:2204.10020.

[12]. Wei, S., Zou, S., & Liao, F. (2020). A comparison on data augmentation methods based on deep learning for audio classification. In Journal of physics: Conference series (Vol. 1453, No. 1, p. 012085). IOP Publishing.

[13]. Seibold, M., Hoch, A., Farshad, M., Navab, N., & Fürnstahl, P. (2022, September). Conditional generative data augmentation for clinical audio datasets. In International Conference on Medical Image Computing and Computer-Assisted Intervention (pp. 345-354). Cham: Springer Nature Switzerland.

[14]. Yun, D., & Choi, S. H. (2022). Deep learning-based estimation of reverberant environment for audio data augmentation. Sensors, 22(2), 592.

[15]. Ko, T., Peddinti, V., Povey, D., Seltzer, M. L., & Khudanpur, S. (2017, March). A study on data augmentation of reverberant speech for robust speech recognition. In 2017 IEEE international conference on acoustics, speech and signal processing (ICASSP) (pp. 5220-5224). IEEE.

[16]. Javed, T., Joshi, S., Nagarajan, V., Sundaresan, S., Nawale, J., Raman, A., ... & Khapra, M. M. (2023). Svarah: Evaluating english ASR systems on indian accents. arXiv preprint arXiv:2305.15760.

[17]. Singh Chadha, H., Gupta, A., Shah, P.,

Chhimwal, N., Dhuriya, A., Gaur, R., & Raghavan, V. (2022). Vakyansh: ASR Toolkit for Low Resource Indic languages. arXiv e-prints, arXiv-2203.

[18]. Palanichamy, H., & Pichai, S. (2025, April). Optimized Voice Activity Detection for Audio Signal Processing. In 2025 3rd International Conference on Advancement in Computation & Computer Technologies (InCACCT) (pp. 726-731). IEEE.

[19]. Arif, S., Khan, A. J., Abbas, M., Raza, A. A., & Athar, A. (2025, January). WER we stand: Benchmarking Urdu ASR models. In Proceedings of the 31st International Conference on Computational Linguistics (pp. 5952-5961).