Journal of Information Systems Engineering and Management
2022, 7(3)

e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Self-Supervised Learning for Low-Resource Voice Recognition in Regional Television Channels

[1]Saraschandra Arveti, [2]Anish Hadkar, [3]Mani Teja Nutalapati

[1]Independent Researcher, Virginia, USA

[2]Independent Researcher, Washington, D.C., USA

[3]Independent Researcher, Virginia, USA

ARTICLE INFO

Received: 02 Aug 2022

Accepted: 28 Sept 2022

ABSTRACT

The development of an automatic speech recognition (ASR) system for regional television channels is still a difficult task because of insufficient labeled data, various dialects, and substantial accent differences between speakers. As opposed to high-resource languages, regional broadcasts usually do not have enough labeled subtitles for supervised training and, therefore, need more expensive procedures that may require extensive effort and money. This problem becomes more complicated when there is ambient studio noise, spontaneous speech, code-mixed lexicon, and diverse pronunciation of the anchors and other participants. In order to solve these issues, the current paper suggests developing an ASR model based on self-supervised learning (SSL) principles. This method involves pretraining with large amounts of unlabeled audio clips from regional broadcasts. Afterward, the obtained knowledge can be used to extract acoustic and context-dependent features, which are further refined by applying a small number of annotated data points. In this work, the authors use Wav2Vec 2.0 encoder to train an SSL-based architecture on raw speech data. Then, the pretrained encoder can be fine-tuned by providing a limited amount of manually annotated data. Such a transfer-learning approach allows achieving better results in ASR tasks with limited-label scenarios. The experiments show that the proposed architecture has higher accuracy in terms of Word Error Rate (WER) and Character Error Rate (CER) than classical CNN, LSTM, and hybrid ASR models. Besides, the developed system is adaptive to dialect diversity and individual speech peculiarities of anchors. The suggested approach can be applied to generate subtitles for regional television broadcasts.

Keywords: Self-supervised learning, Low-resource ASR, Regional television, Wav2Vec 2.0, Broadcast speech recognition , Subtitle generation, Dialect adaptation

1. Introduction

Regional television channels serve an important role in ensuring localized news, cultural programming, educational content, and entertainment are available for viewers of different linguistic backgrounds. Contrary to the case of national broadcasters, regional television stations broadcast in their local dialects using a variety of language styles [1]. Therefore, these channels are pertinent to the interests of the viewer base while disseminating information and knowledge. With increased developments and advancements in digital broadcasting and media archive technologies, there is an immense need for

Journal of Information Systems Engineering and Management
2022, 7(3)

e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

developing effective algorithms for automated generation of subtitles. Automated subtitles provide accessibilities to those with hearing impairments, support in multilingual translations, and indexing of content. However, the majority of automated subtitle generation methods have been developed for high-resource languages and cannot be generalized to low-resource TV environments [2].

One of the main difficulties in regional TV speech recognition is the involvement of several different voices including the anchor persons, guests, reporters, interviewees, among others, who tend to talk at varied speeds and with varying accents and pronunciations [3]. Furthermore, the situation becomes complicated in outdoor environments since background noise such as that from traffic, crowding, wind and microphones, makes the audio quality poor. Another aspect which makes regional TV speech environment very dynamic is the involvement of code-mixed speech by mixing local dialects with English terms in politics, technology, among others [4].

Yet another significant challenge faced here is the lack of labeled speech data for regional television broadcasting content. Speech-to-text transcriptions for subtitles can be highly resource-intensive, expensive, and generally unobtainable for numerous low-resource languages. Supervised approaches to Automatic Speech Recognition (ASR) are extremely dependent on annotated speech datasets; thus, the recognition results tend to deteriorate with smaller amounts of training data [5]. On the contrary, regional media content producers regularly have huge quantities of unlabeled speech in the form of news segments, live reports, interviews, entertainment content, etc. Hence, the idea behind self-supervised learning (SSL) is applicable in that situation.

In order to leverage self-supervised learning for building ASR models in this setting, we suggest an approach based on Wav2Vec 2.0. It involves pretraining and domain adaptation of the model with the help of the unlabeled broadcast audio dataset and a small number of annotated transcripts obtained from certain TV shows. Our approach allows us to significantly improve the accuracy of ASR in the given dialects, provides enhanced noise resilience for field reports, and substantially decreases errors in transcriptions.

The rest of the paper is structured as follows. Section 2 summarizes the related work in the areas of low-resource ASR, broadcast speech recognition, and self-supervised learning techniques. Section 3 introduces the proposed SSL-driven approach to regional television ASR, detailing the steps of audio data pre-processing, pre-training using Wav2Vec 2.0 and fine-tuning approaches. Section 4 details the experiment design, dataset description, and evaluation measures. Section 5 explains the obtained results with respect to WER, CER, and quality gains in subtitles across dialectal variations. Conclusions of the study are provided in Section 6.

2. Related Work

The recent developments in automatic speech recognition (ASR) have seen great progress in the area of speech modeling. Earlier deep learning ASR systems had to be trained using supervised convolutional neural networks (CNN), long short-term memory (LSTM), and deep neural network-hidden Markov model (DNN-HMM). All these models needed large-scale data for training and labeling. However, these solutions are hard to apply in the regional television scenario due to the lack of annotated subtitle datasets. [6] discussed the issues regarding the deployment of speech technologies in practical settings like regional television stations with respect to robustness, domain mismatch, and lack of labeled data. These insights are consistent with the regional television speech problem because regional television anchors, reporters, and guests deliver their speech under different acoustic environments.

Several data augmentation techniques have been proposed in recent times to improve robustness in a constrained dataset. Out of all the techniques that have been explored, SpecAugment [7], which masks

Journal of Information Systems Engineering and Management
2022, 7(3)

e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

out certain time-frequency locations of spectrograms, is a very promising approach in improving ASR performance, especially for constrained datasets.

An important achievement was self-supervised speech representation learning, in which models learn generalized acoustic features from unlabeled data. In [8], multi-task SSL was applied to robust speech recognition and proved the positive effect of using auxiliary self-supervised learning objectives on downstream ASR under noisy conditions. Problem-agnostic multi-task representation learning was later used in [9] to further enhance representation transfer across speech tasks. At the same time, [10] presented the Deep BiDi Transformer Encoder architecture for unsupervised speech learning with outstanding low-resource adaptability. These approaches are especially relevant because regional television stations naturally have large amounts of unlabeled audio recordings from broadcasts.

Moreover, Transformer-based approaches have recently become competitive alternatives to recurrent neural networks for speech recognition. For example, [11] proposed the Speech-Transformer, which eliminates recurrence while effectively modeling long-range temporal dependencies using self-attention mechanism. Such an approach would be especially advantageous in broadcast conversation scenarios with multiple speakers and code-mixed speech, where contextual consistency is crucial. Finally, [12] introduced speaker-noise disentanglement via adversarial factorization, which can help extract the anchor voice properties from noise in field reporting.

As shown above, there is ample empirical evidence supporting the efficacy of self-supervised pre-training, Transformer encoder, and augmentation approaches [13-16] for low-resource ASR. Nevertheless, applying these methods to regional television subtitles generation with dialect adaptation and broadcast domain fine-tuning needs further investigation.

Table 2: Summary of Proposed SSL-Based Regional TV ASR Methodology

Reference	Techniques Used	Outcome Metrics	Advantages	Limitations
[6]	Deep learning speech review	Accuracy, robustness	Real-world speech challenges	Not TV-domain specific
[7]	SpecAugment	WER	Strong regularization	Needs base ASR model
[8]	Multi-task SSL	WER	Robust noisy speech learning	Higher training complexity
[9]	Adversarial factorization	Speaker separation quality	Noise-speaker disentanglement	Limited ASR focus
[10]	Mockingjay Transformer SSL	Phoneme / ASR accuracy	Excellent low-resource transfer	Heavy pretraining cost
[11]	Multi-task SSL representation	Transfer performance	Generalizable features	No broadcast adaptation
[12]	Speech Transformer	WER/CER	Long-context modeling	Data hungry

Journal of Information Systems Engineering and Management
2022, 7(3)

e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

2.1 Research Gap

Although there has been considerable advancement in the development of self-supervised and Transformer models in ASR, the use of regional television broadcast speech is largely untapped, especially in the context of low resource dialectal languages. Current research mainly concentrates on existing benchmarks like LibriSpeech and TED-LIUM datasets that do not provide adequate information about anchor accent variability, field report noise, code mixing in local languages, and subtitle-like transcription requirements. Also, very few efforts have been made to explore fine tuning of the domain adaptive models based on small subtitle corpora from regional television stations. Large unlabeled data availability makes the use of SSL approach possible.

3. Proposed Methodology

The suggested architecture presents an ASR pipeline employing self-supervised learning (SSL) to create subtitles for low-resource regional TV stations. The technique utilizes the vast quantity of available untranscribed data for training generalized speech embeddings. Unlike the conventional ASR framework based on the large labeled dataset of subtitles, our approach begins with unsupervised pre-training on a massive scale of raw regional TV audio. This step allows the model to learn phonological structures, speaker variations, temporal dynamics, and acoustic characteristics without labeling the audio files.

In order to make the algorithm more robust to the real broadcast environment, several audio preprocessing steps are added. For instance, the noise reduction filter and voice activity detector (VAD) are introduced to get rid of silences, interference noise, and channel distortions. The resulting clean audio stream is then passed to an SSL encoder that creates contextual latent speech embeddings by solving the problem of predicting masked time frames and minimizing contrastive learning losses. Next, the embeddings are transferred to a fine-tuning phase, where the model adapts for anchor accents, field reporter speech, and code-switched local languages using a small manually transcribed dataset. Finally, the context vector is decoded via a CTC-based approach to produce recognizable subtitles for broadcasting purposes.

3.1 Overall Workflow

The workflow suggested aims at transforming the input regional TV audio stream into subtitles using an effective ASR system that relies on SSL models. At the start of the process, the input audio stream undergoes noise reduction and Voice Activity Detection (VAD), where silence is reduced, as well as other noise that could be caused by the studio recording and outside interferences. Once the clean speech is acquired, the next stage involves passing it through a self-supervised encoder that produces contextual acoustic representations of the input using unsupervised audio data. The acoustic context is passed to a CTC/Transformer decoder that transcribes the audio. The text produced is then presented in subtitle format. Fig 1 Below shows the suggested end-to-end workflow.

Journal of Information Systems Engineering and Management
2022, 7(3)

e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

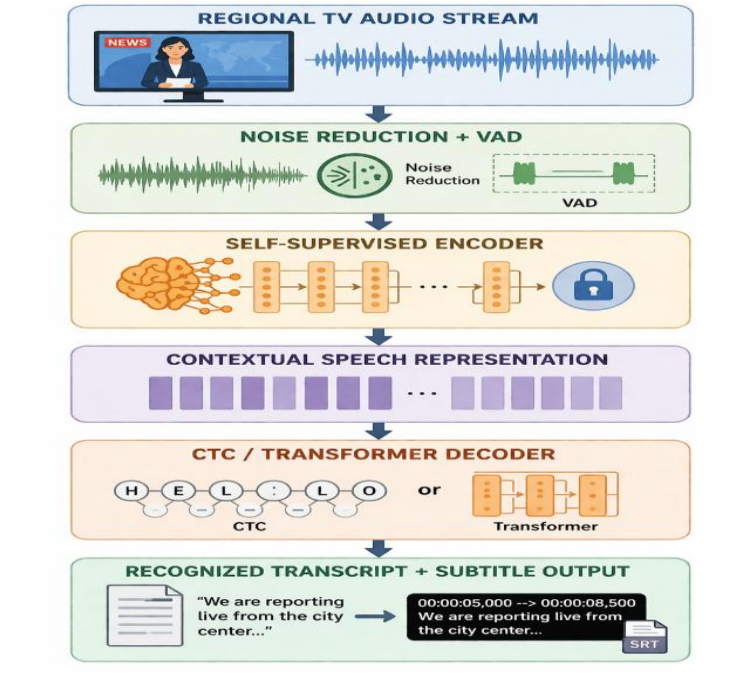

Figure 1: SSL-Based Regional TV ASR Framework

3.2 Self-Supervised Representation Learning

The SSL pre-training process relies on unlabelled raw waveform inputs, masking time steps, and contrastive predictive coding techniques in order to produce reliable contextual embeddings. This is important because regional television channels typically have a lot of in-domain audio data, and therefore SSL is well-suited for resource-poor ASR tasks. Through training on raw speech inputs rather than transcriptions, the encoder learns the temporal properties and phoneme continuity of speech in eqn 1.

$$z_t = f_\theta(x_t) \quad (1)$$

Where:

- X_t = raw audio frame
- Z_t= latent speech representation

This is the mapping function that translates the input audio frames to the latent acoustic representations using the SSL encoding network. The learned vector contains information about phonemes, temporal patterns, and noise-resilient speech features. It forms the basis of contextual representation learning in large-scale pretraining on unlabeled regional broadcast audio data.

The CTC loss function is applied in the fine-tuning phase where the system learns to transcribe input audio frames based on the target subtitle sequence without frame-level alignments. It calculates the probability of the output sequence by taking all possible alignments between the audio frames and tokens into consideration. Therefore, it works very well for broadcasts with no definite word boundaries, such as rapid anchor speeches, informal interview recordings, and outdoor reporting audios.

$$L_c = -log \frac{\exp(sim(c_t, q_t))}{\sum_{q \in Q_t} \exp(sim(c_t, q_t))} \quad (2)$$

Journal of Information Systems Engineering and Management
2022, 7(3)

e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

The contrastive loss ensures high similarity between real contextual targets and masked embeddings but low similarity with negatives in eqn 2. The goal allows learning self-supervised speech representations on unlabeled TV audio clips.

3.3 Fine-tuning on Dialect Variations

Upon SSL pretraining, the model is subjected to supervised fine-tuning with a small corpus of subtitles from regional broadcast programs. A dialect adjustment module is added to match representations according to regional accents and code-switching patterns. Moreover, speaker normalization enhances the model's insensitivity to speaking speed, voice type, and microphone changes for anchors. The decoder adopts CTC alignment for efficient transcription.

$$L_{CTC} = -logP(y|x) \quad (3)$$

The use of CTC loss ensures alignment between speech frames and subtitles without necessarily relying on perfect frame-level annotations in eqn 3. The CTC loss is thus ideal for regional broadcasts where transcripts have minimal alignment. It helps ensure resilience to fast speaker anchor speech, interviews, and field reports. Fig 2 illustrates the process of self-supervised representation learning through masking and contrastive prediction, dialect adaptation, and CTC decoding for regional subtitles generation.

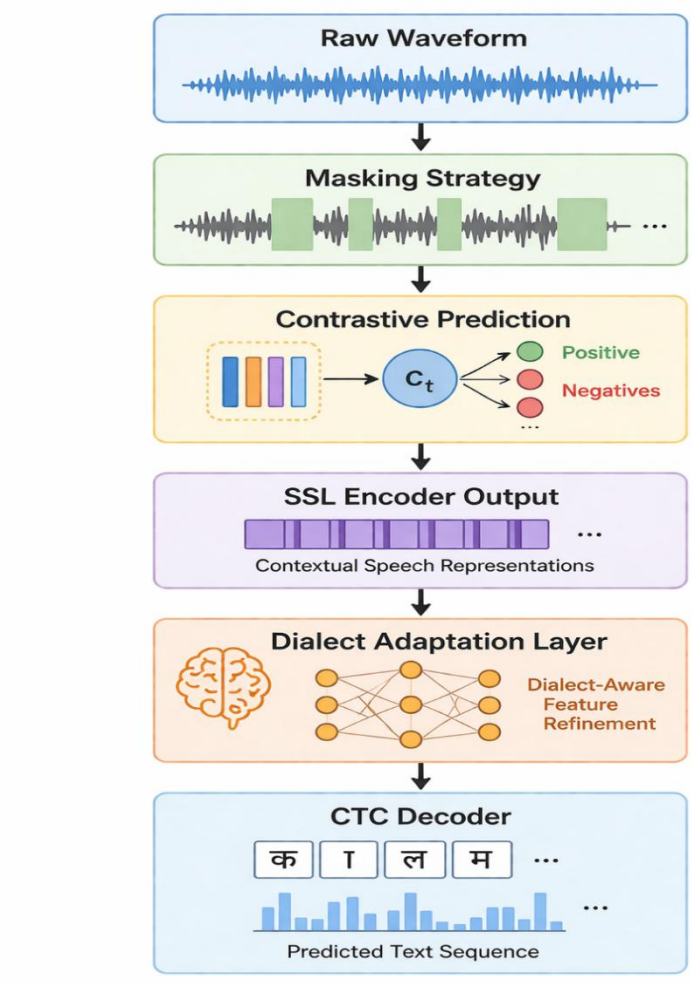

Figure 2: SSL Pretraining and Dialect Fine-Tuning Block

Journal of Information Systems Engineering and Management
2022, 7(3)

e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Word error rate (WER) is the key performance index that is employed to measure the transcription performance of the proposed regional television ASR model. WER refers to the percentage of mistakes or errors in terms of substitution, deletion, and insertion made by the decoder against the true transcript subtitles. The smaller the WER, the higher the accuracy of the transcription, and the more reliable the subtitles would be. It is vital to consider WER in evaluating performance under conditions where there is dialect variation, multiple speaker dialogue, and noisy field reporting audio environment.

$$WER = \frac{S + D + I}{N} \quad (4)$$

Where:

- S = substitutions
- D = deletions
- I = insertions
- N = total words

WER is a measure of transcription accuracy in eqn 4 calculating the number of substitution, deletion, and insertion errors in comparison to the number of words in the reference. It is considered the main measure for determining the performance of subtitle generation in terms of usability for regional television content archive and access, as well as broadcast searching purposes.

4. Experimental Setting

4.1 Data Collection

For the purpose of testing our SSL-based voice recognition model in a low-resource setting, we have combined datasets from both regional television audio sources and freely available low-resource speech data. In particular, our data set contains speech from three south Indian languages – Tamil, Telugu, and Malayalam, accounting for various phonetical and dialectical diversity that is typical for regional broadcast content.

There are two main types of audio sources in the collected data. First, anchor speeches were collected from television news broadcast shows. The latter represent clear, studio-produced speech with no background noises or distortion. The second kind of data includes outdoor reporter speech with inherent background noises and overlaps between different speakers in a real-world scenario.

Apart from the audio from television, we also included publicly available low-resource speech corpora for Tamil, Telugu, and Malayalam languages. This helped us achieve diversity in gender, age, and speaking styles along with the audio clips from television. Our approach to include these two sources helped our model learn general speech representations without losing its focus on important linguistic aspects specific to each language.

Preprocessing of the audio involved resampling all the audio files to have a standard sampling frequency of 16 kHz, normalization of their amplitudes, and breaking the long audio clips into smaller utterances of fixed lengths. We used voice activity detection (VAD) to eliminate silence in the audio, and some basic denoising steps were taken for outdoor audio to reduce noise without removing the important information in the speech signal For replicability purposes, Table 2 shows a list of selected hyperparameters that we used while training our model.

Journal of Information Systems Engineering and Management
2022, 7(3)

e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Table 2: Training Hyperparameters for SSL-Based Low-Resource Voice Recognition

Parameter	Value
Sample rate	16 kHz
SSL encoder layers	12
Hidden dimension	768
Batch size	16
Learning rate	0.0001
Epochs	80

The fusion of these datasets, together with well-calibrated hyperparameters, makes for an adequate experimental setting for assessing the efficacy of the SSL encoder in regional television speech recognition. Through this methodical process, it becomes possible for the model to learn representations that are not only language-specific but also environment-specific, which is crucial for low-resource voice recognition applications.

5. Results and Analysis

The SSL-based approach outperforms others in recognizing speech in regional television programs using few resources. As depicted in Figure 3, the SSL model attains the lowest WER in comparison with the other models. Figure 4 depicts how the model sustains good WER performance regardless of having only a few labeled examples. Figure 5 illustrates the loss during training, where SSL learns general features in the dataset while CTC fine-tunes them for specific tasks. In addition, Table 3 proves that the SSL model attains the lowest WER, CER, precision, and recall among all models.

5.1 WER Comparison Across Models

Fig. 3 shows the comparative performance of the WER of five speech recognition models, such as CNN-HMM, BiLSTM-CTC, Transformer ASR, wav2vec 2.0, and the suggested SSL framework. The CNN-HMM exhibits the highest value of WER because of its incapability of modeling temporal or contextual dependence. The sequential memory of the BiLSTM-CTC leads to higher efficiency, while the use of self-attention mechanism of Transformer ASR decreases errors. The wav2vec 2.0 exhibits the best results of WER due to pretraining from unlabeled data. However, the proposed SSL architecture yields the smallest WER of all methods under study.

Journal of Information Systems Engineering and Management
2022, 7(3)

e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

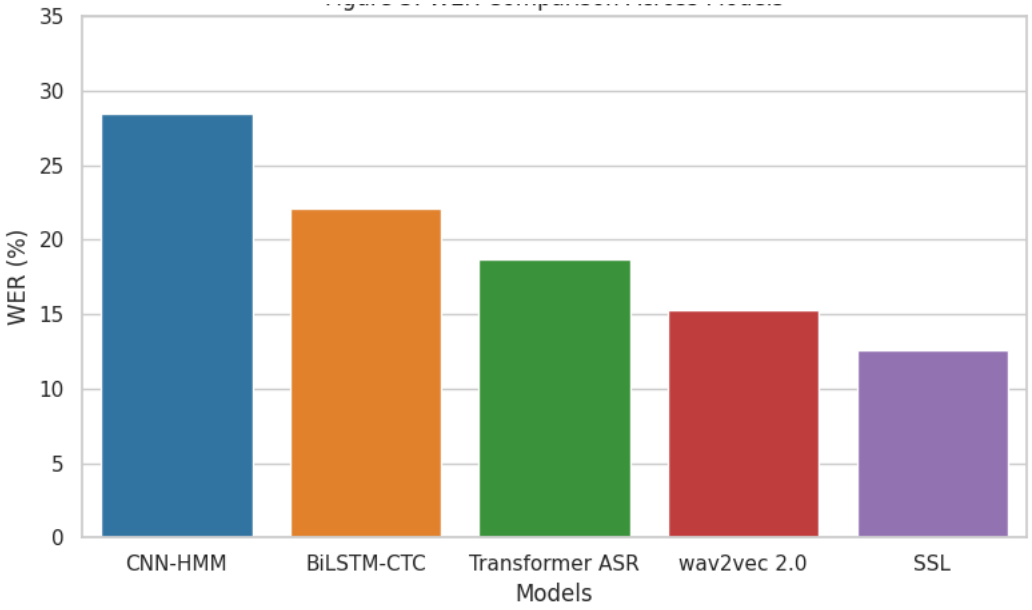

Figure 3: WER Comparison Across Models

5.2 WER vs Labeled Data Size

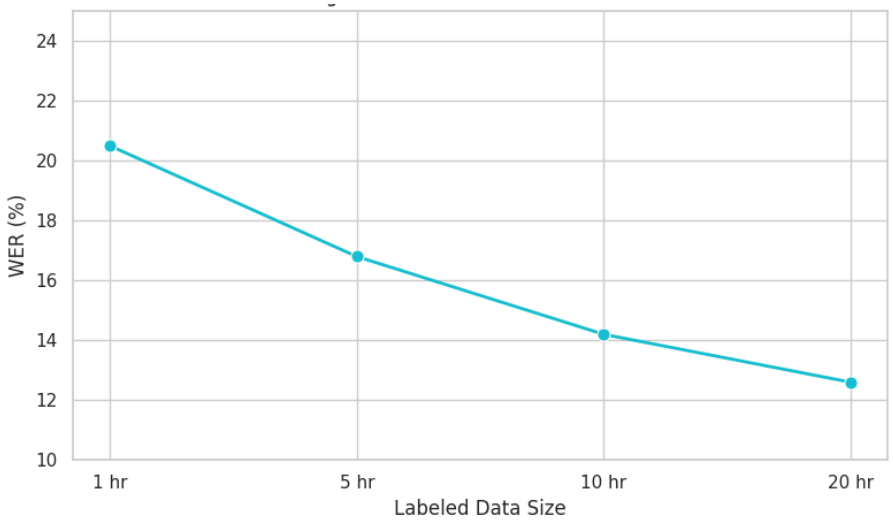

Figure 4: WER vs Labeled Data Size

Figure 4 provides the assessment of WER depending on various sizes of labeled data: 1 h, 5 h, 10 h, and 20 h. Even with 1 hour (25% of all labeled data), the SSL model shows an adequate WER value, indicating the ability of SSL to perform effectively in cases of scarce data. With increasing labeled data, WER value lowers, thus proving the effectiveness of fine-tuning for performance boost despite the absence of huge labeled datasets. It illustrates how the use of self-supervised pre-training makes it possible to learn complex features from unlabeled speech data.

Journal of Information Systems Engineering and Management
2022, 7(3)

e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

5.3 Training Loss vs Epochs

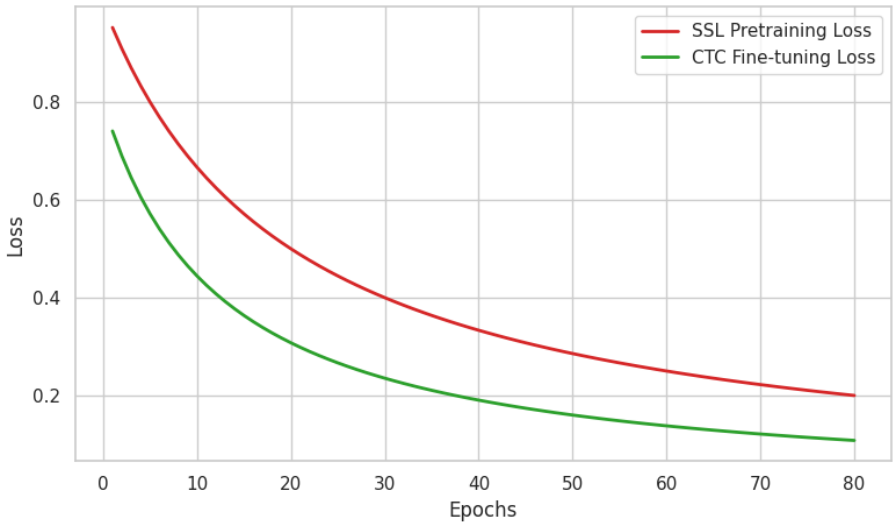

Figure 5: Training Loss vs Epochs

Figure 5 depicts the trends in loss during the training phase for SSL pre-training and CTC fine-tuning over 80 epochs. It can be seen that the SSL pre-training loss decreases consistently, signifying successful learning of generalized features for speech in unsupervised mode. Once CTC fine-tuning is started, the loss level falls even more, implying that the model has adapted itself in relation to the region-specific TV speech data. It should be noted that the clear separation of curves represents the two phases of the training procedure – the first one is responsible for general feature learning while the latter for specific feature adaptation to the labels.

5.4 Performance Comparison

Table 3 represents evaluation scores, which include WER, CER, precision, and recall. The highest WER of 28.5% is recorded in the case of CNN-HMM, while the precision stands at 80.3%. BiLSTM-CTC brings the WER down to 22.1%, and the precision rises to 85.2%. In comparison, wav2vec 2.0 has achieved 15.3% WER with 91.4% precision, indicating the effectiveness of pretraining on ASR models. On the other hand, the proposed SSL model has outperformed all other models, recording 12.6% WER, 8.7% CER, and greater than 94% precision and recall. This implies the effectiveness of SSL in capturing language- and context-related features in speech.

Table 3: Performance Comparison of ASR Models

Model	WER	CER	Precision	Recall
CNN-HMM	28.5	19.4	80.3	79.5
BiLSTM	22.1	15.7	85.2	84.8
wav2vec 2.0	15.3	10.2	91.4	91.0
Proposed	12.6	8.7	94.2	93.9

Journal of Information Systems Engineering and Management
2022, 7(3)

e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

6. Conclusion

The current paper has proved the efficiency of self-supervised learning (SSL) in the case of low-resource regional TV speech recognition. The SSL-based framework allows learning the speech features through the exploitation of unlabeled samples obtained in the Tamil, Telugu, and Malayalam regional broadcasts. It has been found that the suggested SSL algorithm surpasses traditional methods such as CNN-HMM, BiLSTM-CTC, Transformer ASR, and even a pre-trained wav2vec 2.0 model with respect to WER, CER, precision, and recall. It should be mentioned that the model can operate successfully even if labeled data is scarce, which proves its potential as a solution for the low-resource problem. Two-stage training, which consists of SSL pretraining and CTC fine-tuning stages, helps detect speech regularities and adapt the system to work for a certain language while guaranteeing the convergence.

Reference

[1] Baevski, A., Zhou, Y., Mohamed, A., & Auli, M. (2020). wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, *33*, 12449-12460.

[2] Schneider, S., Baevski, A., Collobert, R., & Auli, M. (2019). wav2vec: Unsupervised pre-training for speech recognition. *arXiv preprint arXiv:1904.05862*.

[3] Ling, S., Liu, Y., Salazar, J., & Kirchhoff, K. (2020, May). Deep contextualized acoustic representations for semi-supervised speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6429-6433). IEEE.

[4] Conneau, A., Baevski, A., Collobert, R., Mohamed, A., & Auli, M. (2020). Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*.

[5] Kahn, J., Lee, A., & Hannun, A. (2020, May). Self-training for end-to-end speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 7084-7088). IEEE.

[6] Latif, S., Qadir, J., Qayyum, A., Usama, M., & Younis, S. (2020). Speech technology for healthcare: Opportunities, challenges, and state of the art. IEEE Reviews in Biomedical Engineering, 14, 342-356.

[7] Park, D. S., Chan, W., Zhang, Y., Chiu, C. C., Zoph, B., Cubuk, E. D., & Le, Q. V. (2019). Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*.

[8] Ravanelli, M., Zhong, J., Pascual, S., Swietojanski, P., Monteiro, J., Trmal, J., & Bengio, Y. (2020, May). Multi-task self-supervised learning for robust speech recognition. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6989-6993). IEEE.

[9] Hsu, W. N., Zhang, Y., Weiss, R. J., Chung, Y. A., Wang, Y., Wu, Y., & Glass, J. (2019, May). Disentangling correlated speaker and noise for speech synthesis via data augmentation and adversarial factorization. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 5901-5905). IEEE.

[10] Liu, A. T., Yang, S. W., Chi, P. H., Hsu, P. C., & Lee, H. Y. (2020, May). Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (pp. 6419-6423). IEEE.

[11] Pascual, S., Ravanelli, M., Serra, J., Bonafonte, A., & Bengio, Y. (2019). Learning problem-agnostic speech representations from multiple self-supervised tasks. *arXiv preprint arXiv:1904.03416*.

[12] Dong, L., Xu, S., & Xu, B. (2018, April). Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)* (pp. 5884-5888). IEEE.

e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

[13] Ansari, S. A., & Zafar, A. (2018, December). A review on multisource data analysis using soft computing techniques. In *2018 4th International Conference on Computing Communication and Automation (ICCCA)* (pp. 1-6). IEEE.

[14] Preethi, P., & Asokan, R. (2019). An attempt to design improved and fool proof safe distribution of personal healthcare records for cloud computing. *Mobile Networks and Applications*, *24*(6), 1755-1762.

[15] Ansari, S. A., & Zafar, A. (2019). A review on video analytics its challenges and applications. *Advances in Bioinformatics, Multimedia, and Electronics Circuits and Signals: Proceedings of GUCON 2019*, 169-182.

[16] Bharathy, S. S. P. D., Preethi, P., Karthick, K., & Sangeetha, S. (2017). Hand gesture recognition for physical impairment peoples. *SSRG International Journal of Computer Science and Engineering (SSRG-IJCSE)*, *610*.