How to Cite:
Kumar, L. A., Renuka, D. K., Shunmuga, P. M. C., Madhumitha, G., Priyanka, S., Sangeeth, M., & Subhiksha, R. (2022). Self-supervised learning based knowledge distillation framework for automatic speech recognition for hearing impaired. *International Journal of Health Sciences*, *6*(S1), 11728–11737. https://doi.org/10.53730/ijhs.v6nS1.7865

Self-supervised learning based knowledge distillation framework for automatic speech recognition for hearing impaired

L. Ashok Kumar
Professor, Dept. of EEE, PSG College of Technology

D. Karthika Renuka
Associate Professor, Dept. of IT, PSG College of Technology

Shunmuga Priya M C
Research Scholar, Dept. of IT, PSG College of Technology

Madhumitha G
UG Students, Dept. of IT, PSG College of Technology

Priyanka S
UG Students, Dept. of IT, PSG College of Technology

Sangeeth M
UG Students, Dept. of IT, PSG College of Technology

Subhiksha R
UG Students, Dept. of IT, PSG College of Technology

***Abstract*---**The use of speech processing applications, particularly speech recognition, has got a lot of attention in recent decades. In recent years, research has focused on using deep learning for speech-related applications. This new branch of machine learning has outperformed others in a range of applications, including voice, and has thus become a particularly appealing research subject. Noise, speaker variability, language variability, vocabulary size, and domain remain one of the most significant research difficulties in speech recognition. We investigated on self-supervised algorithm for the unlabelled data. In recent years, these algorithms have progressed significantly, with their efficacy approaching and supervised pre-training alternatives across a variety of data modalities such as image and video. The purpose of this research is to develop powerful models for audio speech recognition that do not require human annotation. We accomplish this by distilling information from an automatic speech

recognition (ASR) model that was trained on a large audio-only corpus. We integrate Connectionist Temporal Classification (CTC) loss, KL divergence loss in distillation technique. We demonstrate that distillation significantly speeds up training. We evaluate our model with evaluation metric Word Error Rate (WER).

***Keywords*---**Automatic Speech Recognition, Self-Supervised, Knowledge Distillation, WER, Deep Learning.

Introduction

Speech has been an important means of human communication for thousands of years, they communicate with each other in many ways such as speech, hand gestures, facial expressions… etc. But speech is considered the most important means that a human uses, as it facilitates communication and it is the most widely used between speakers. Speech is a useful expression and has a particular meaning and it is composed of several words, which in turn contain several letters accompanied by voices. This voice can spread in objects of air and empty and inanimate in the form of waves; a wave that overlaps between them or begins in the form of small circles of the sound source. The process of turning acoustic speech into text and identifying the speaker is known as speech recognition. Due to the increasing contact between humans and computers or automated systems, technology has become a vital and integral element of our existence over the years. Types of Speech Recognition Systems are some of the most extensively used speech recognition systems. Speaker Dependent Systems, Speaker Independent Systems, Isolated Word Recognizers, Connected Word Recognizers, and Spontaneous Recognition Systems are some examples of speaker dependent systems. The majority of previous study has been attributed to the fact that speech is a highly subjective process. Self-supervised learning has emerged as a paradigm to be told cognition representations from unlabelled examples and to fine-tune the model on labelled knowledge. Deep learning algorithms have primarily been used to improve the skills of computers so that they can grasp what humans can accomplish, such as speech recognition. Speech, as the primary mode of human communication, has peaked researchers' attention for the past five decades, going back to the development of artificial intelligence. As a result, it's only logical that speech was one of the first uses of deep learning. This has been notably triple-crown for language method. Here we tend to tend to gift a framework for self-supervised learning of representations from raw audio knowledge. Our approach encodes speech audio via a multi-layer convolutional neural network and then masks spans of the resulting latent speech representations.

Self-supervised learning has emerged as a paradigm to be told cognition representations from unlabelled examples and to fine-tune the model on labelled knowledge. This has been notably triple-crown for language method. Here we tend to tend to gift a framework for self-supervised learning of representations from raw audio knowledge. Our approach encodes speech audio via a multi-layer convolutional neural network then masks spans of the following latent speech representations, like disguised language modelling. This very promising approach

is that we tend to identify the knowledge in a trained model with the learned parameter values and this makes it hard to see how we can change the form of the model but keep the same knowledge. A more abstract view of the knowledge, that frees it from any particular instantiation, is that it is a learned mapping from input vectors to output vectors. The goal of knowledge distillation is to extract knowledge from a teacher model into a student model. Typically, the teacher model is a large neural network and the student model is a smaller version with fewer layers. In this direction, we propose to train a VSR model by distilling from an ASR model with a teacher-student approach. Training by distillation eliminates the need for professionally transcribed subtitles, and also removes the Information compresses and moves from the teacher model into the student model through knowledge distillation training. Wav2vec has two major components: CNN layers and transformer layers. Our student wav2vec model has fewer transformer layers than the teacher model, but the same number of CNN layers. It's possible to reduce the number of CNN or transformer layers when designing the student model. Here, we choose not to reduce CNN layers because there are only seven, and they are not a huge computational bottleneck compared to transformer layers.

Related Work
Offline to Online Knowledge Distillation in ASR

Frame-level alignments between audio and output symbols are not required for end-to-end training of recurrent neural network transducers (RNN-Ts). As a result, the posterior lattices generated by the prediction distributions from multiple RNN-Ts trained on the same data can vary significantly, posing additional issues in knowledge distillation between such models. The differences between an offline and a streaming model are most noticeable in the posterior lattices, which is to be expected given that the streaming RNN-T emits symbols later than the offline RNN-T. We offer a method for training an RNN-T so that the posterior peaks at each node in the posterior lattice match those from a previously learned model for the same phrase(Gakuto Kurataty et al., 2020).We can train an offline RNN-T that can act as a good teacher for a student streaming RNN-T using this strategy. Experiments on the standard Switchboard conversational telephone speech corpus show that knowledge distillation from an offline bidirectional equivalent improves accuracy for a streaming unidirectional RNN-T.

Multi-Teacher Distillation

This study looks at the issue from the standpoint of regulating the level of precision with which the teacher network is trained. Existing systems often used a rigid distribution (e.g., one-hot vectors) in training, resulting in a strict teacher with high accuracy, but we suggest that the teacher should be more forgiving, despite the fact that this often implies lesser accuracy. The implementation is simple, requiring only the addition of an extra loss term to the teacher network, allowing for the emergence of a few supplementary courses to supplement the primary class. As a result, the teacher sends out a less peaked supervision signal, allowing the student to benefit from inter-class similarity and potentially reducing the risk of overfitting (Chenglin Yang et al., 2018). Standard picture classification

tasks are used in the experiments. Although the teacher network is less robust, the students' abilities increase over time and they finally attain higher classification accuracy than their competition. The efficiency of our approach is also shown by model ensemble and transfer feature extraction.

Table 1
Literature Survey

Model Type	Papers Title	Inference
Offline to Online KD in ASR	Knowledge Distillation from Offline to Streaming RNN Transducer for End-to-end Speech Recognition[1.1]	This paper provides us with method to train RNN-T that can serve as a good teacher to train a student streaming RNN-T
	Dual-Mode Asr: Unify And Improve Streaming Asr With Full-Context Modeling[1.2]	This paper proposes Dual-mode ASR, a framework to unify streaming and full-context speech recognition networks with shared weights
Cross Modal KD	Asr Is All You Need: Cross-Modal Distillation For Lip Reading[2.1]	This paper proposes a method to train a VSR model by distilling from an ASR model with a teacher-student approach.
	Um-Adapt: Unsupervised Multi-Task Adaptation Using Adversarial Cross-Task Distillation[2.2]	This paper proposes two novel regularization strategies; a) Contour-based content regularization (CCR) and b) exploitation of inter-task coherency using a cross-task distillation module.
	Learning Deep Representations With Probabilistic Knowledge Transfer[2.3]	This paper proposes a Probabilistic KT (PKT) technique that overcomes several limitations of existing KT methods by matching the probability distribution of the data in the feature space instead of their actual representation.
	Knowledge As Priors: Cross-Modal Knowledge Generalization For Datasets Without Superior Knowledge[2.4]	This paper proposes a technique to transfer learned cross-modal knowledge from a source dataset, where both modalities are available, to the target dataset, where only one weak modality exists.
	Cross-Modal Knowledge Distillation For Action Recognition[2.5]	This paper proposes the crossentropy for the transfer from the teacher to the student and train not one student network, but multiple student networks.
Large model to fine grained one	Structural Knowledge Distillation: Tractably Distilling Information For Structured Predictor[3.1]	This paper provides us with a method to derive a factorized form of the knowledge distillation objective for structured prediction, which is tractable for wide range of typical choices of the teacher and student models.
Multi-teacher Distillation	Training Deep Neural Networks In Generations: A More Tolerant Teacher Educates Better	This paper provides us with an efficient “tolerant-teacher” framework which achieves superior performance.

	Students[4.1]	
	Model Compression With Two-Stage Multi-Teacher Knowledge Distillation For Web Question Answering System[4.2]	This paper proposes a Two-stage Multi-teacher Knowledge Distillation (TMKD for short) method for model compression
	Feature-Level Ensemble Knowledge Distillation For Aggregating Knowledge From Multiple Networks[4.3]	This paper proposes parallel FEED, a method that allows multiple teacher networks to be used for knowledge distillation
	Stochasticity And Skip Connection Improve Knowledge Transfer[4.4]	This paper proposes to add stochastic blocks and skip connections to a teacher network
	Born-Again Neural Networks[4.5]	This paper proposes to revisit KD with the objective of disentangling the benefits of this training technique from its use in model compression.
Adversarial Distillation	Improved Knowledge Distillation Via Teacher Assistant[5.1]	This paper proposes a new distillation framework called Teacher Assistant Knowledge Distillation (TAKD), which introduces intermediate models as teacher assistants (TAs) between the teacher and the student
	Zero-Shot Knowledge Transfer Via Adversarial Belief Matching[5.2]	This paper proposes a novel adversarial algorithm that distills a large teacher into a smaller student without any data or metadata
	Ktan: Knowledge Transfer Adversarial Network[5.3]	This paper proposes to exploit intermediate representations as sharable knowledge. Specifically, they use the outputs in the convolutional layers of a teacher network.
	Data-Free Knowledge Amalgamation Via Group-Stack Dual-GAN[5.4]	This paper proposes a new data-free knowledge amalgamation framework for training the target network
	Feature-Map-Level Online Adversarial Knowledge Distillation[5.5]	This paper proposes an online knowledge distillation method that utilizes not only the logit but also the feature map from the convolution layer

Methodology
Wav2vec:

Our model is composed of a multi-layer convolutional feature encoder f : X 7→ Z which takes as input raw audio X and outputs latent speech representations z1, . . . , zT for T time-steps. They are then fed to a Transformer g : Z 7→ C to build representations c1, . . . , cT capturing information from the entire sequence. The output of the feature encoder is discretized to qt with a quantization module Z 7→

Q to represent the targets in the self-supervised objective. Compared to vq-wav2vec, our model builds context representations over continuous speech representations and self-attention captures dependencies over the entire sequence of latent representations end-to-end.

Feature encoder. The encoder consists of several blocks containing a temporal convolution followed by layer normalization and a GELU activation function. The raw waveform input to the encoder is normalized to zero mean and unit variance. The total stride of the encoder determines the number of time-steps T which are input to the Transformer.

Contextualized representations with Transformers. The output of the feature encoder is fed to a context network which follows the Transformer architecture [55, 9, 33]. Instead of fixed positional embeddings which encode absolute positional information, we use a convolutional layer similar to [37, 4, 57] which acts as relative positional embedding. We add the output of the convolution followed by a GELU to the inputs and then apply layer normalization.

Quantization module. For self-supervised training we discretize the output of the feature encoder z to a finite set of speech representations via product quantization. This choice led to good results in prior work which learned discrete units in a first step followed by learning contextualized representations.

Training - To train the model we mask a certain proportion of time steps in the latent feature encoder space, similar to masked language modelling in BERT.

Masking - We mask a proportion of the feature encoder outputs, or time steps before feeding them to the context network and replace them with a trained feature vector. To mask the latent speech representations output by the encoder, we randomly sample without replacement a certain proportion p of all time steps to be starting indices and then mask the subsequent M consecutive time steps from every sampled index.

Loss - The objective of pre-training could be a total of 2 loss functions: contrastive loss is chargeable for coaching the model to predict and variety loss it permits the network to come up with various texture samples. Models are optimized by minimizing a CTC loss and that we apply a changed version of Spec Augment by masking to time-steps and channels that improves the ultimate error rates ,especially on the Libri-light subsets with few tagged examples.

Fine-tuning - Pre-trained models are fine-tuned for speech recognition by adding a randomly initialized linear projection on top of the context network into C classes representing the vocabulary of the task. The proposed architecture shown in figure 1.

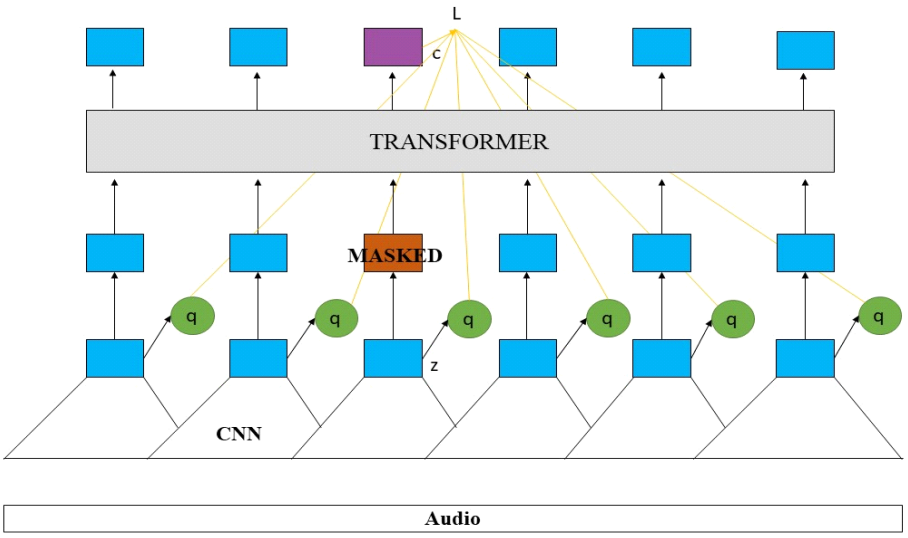

Fig 1: Wav2vec Architecture

Knowledge Distillation

The probability distributions over the possible tokens are output by both the student and teacher models. The Kullback–Leibler (KL) divergence between the probability distributions of the student and teacher models is used to evaluate knowledge distillation loss. We want the probability distribution of the student model to be similar to that of the teacher model. For both the teacher and student models, the probability distribution is computed by taking the output of the transformer layers, passing it to a linear layer, and then performing a softmax operation. To quantify the loss, we must first obtain the probability distributions of the instructor and student models. We enter the input into the instructor model to determine the probability distribution. batch is a tuple that contains raw audio waveforms, padding masks, and other parameters. When the batch size is greater than 1, the input waveform is padded to the length of the batch's longest waveform. The padding mask parameter indicates which part of the input waveform should be masked. Calculate the log likelihood for the student model next, as we'll need it to figure out the knowledge distillation loss later. The student model, unlike the teacher model, is in training mode. The KL divergence between the probability distributions of the teacher model and the student model is the knowledge distillation loss.

Loss - Knowledge distillation loss is measured using the Kullback–Leibler (KL) divergence between probability distributions of the student and the teacher model. We want the student model to have a similar probability distribution to the teacher model. The process of computing the probability distribution is the same for both the teacher and student models: we take the output of the transformer layers, then pass it to a linear layer, followed by a softmax operation. We need to first get the probability distribution of the teacher and the student model, then calculate the loss. The proposed architecture shown in figure 2.

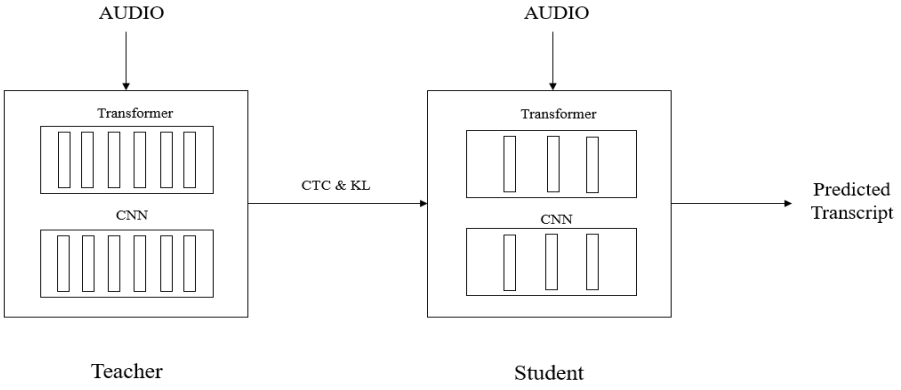

Fig 2: Knowledge Distillation framework

Result Analysis

In this part, we present our experimental results of different configurations. After applying the Knowledge Distillation on the audio and audio module, we again calculated WER for true and predicted transcription obtained from the model. The average WER of the audio module i.e., in the Wav2vec model is 0.0338. After Knowledge Distillation from the audio module to audio module, the average WER is 0.0576.

Table 2
Comparison of Wav2Vec with Baseline models

Model	WER
RNN-GRU	0.6455
RNN-GRU+Bert	0.5236
LAS	0.5540
Wav2Vec	0.0338
ASR to ASR KD	0.0576

Table 3
Performance of Wav2Vec with different labelled and unlabelled data ratio

Labelled data	Unlabelled Data	WER
60	40	0.0763
70	30	0.0471
80	20	0.0338

Conclusion

The Recent work shows that a collaborative distillation framework is employed to share knowledge across different models. Knowledge Distillation can improve the performance of student model and also make student model perform like teacher model. The student model is more suitable for deployment because it will be computationally less expensive than the Teacher model while maintaining the same or better accuracy. Extensive experiments demonstrate that the proposed model can perform better than the existing audio speech models. Moreover, it is worthwhile to notice that this approach is capable of promoting the performance of the model.

References

1. Kurata, G., & Saon, G. (2020). Knowledge Distillation from Offline to Streaming RNN Transducer for End-to-End Speech Recognition. In Interspeech (pp. 2117-2121).
2. Yang, C., Xie, L., Qiao, S., & Yuille, A. L. (2019, July). Training deep neural networks in generations: A more tolerant teacher educates better students. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, No. 01, pp. 5628-5635).
3. Yu, J., Han, W., Gulati, A., Chiu, C. C., Li, B., Sainath, T. N., ... & Pang, R. (2020). Dual-mode asr: Unify and improve streaming asr with full-context modeling. arXiv preprint arXiv:2010.06030.
4. Afouras, T., Chung, J. S., & Zisserman, A. (2020, May). Asr is all you need: Cross-modal distillation for lip reading. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 2143-2147). IEEE.
5. Kundu, J. N., Lakkakula, N., & Babu, R. V. (2019). Um-adapt: Unsupervised multi-task adaptation using adversarial cross-task distillation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 1436-1445).
6. Passalis, N., & Tefas, A. (2018). Learning deep representations with probabilistic knowledge transfer. In Proceedings of the European Conference on Computer Vision (ECCV) (pp. 268-284).
7. Zhao, L., Peng, X., Chen, Y., Kapadia, M., & Metaxas, D. N. (2020). Knowledge as priors: Cross-modal knowledge generalization for datasets without superior knowledge. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 6528-6537).
8. Thoker, F. M., & Gall, J. (2019, September). Cross-modal knowledge distillation for action recognition. In 2019 IEEE International Conference on Image Processing (ICIP) (pp. 6-10). IEEE.
9. Wang, X., Jiang, Y., Yan, Z., Jia, Z., Bach, N., Wang, T., ... & Tu, K. (2020). Structural Knowledge Distillation: Tractably Distilling Information for Structured Predictor. arXiv preprint arXiv:2010.05010.
10. Yang, C., Xie, L., Qiao, S., & Yuille, A. L. (2019, July). Training deep neural networks in generations: A more tolerant teacher educates better students. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 33, No. 01, pp. 5628-5635).

11. Yang, Z., Shou, L., Gong, M., Lin, W., & Jiang, D. (2020, January). Model compression with two-stage multi-teacher knowledge distillation for web question answering system. In Proceedings of the 13th International Conference on Web Search and Data Mining (pp. 690-698).
12. Park, S., & Kwak, N. (2020). Feature-level ensemble knowledge distillation for aggregating knowledge from multiple networks. In ECAI 2020 (pp. 1411-1418). IOS Press.
13. Nguyen, L. T., Lee, K., & Shim, B. (2021, January). Stochasticity and Skip Connection Improve Knowledge Transfer. In 2020 28th European Signal Processing Conference (EUSIPCO) (pp. 1537-1541). IEEE.
14. Vidal, T., & Schiffer, M. (2020, November). Born-again tree ensembles. In International conference on machine learning (pp. 9743-9753). PMLR.
15. Mirzadeh, S. I., Farajtabar, M., Li, A., Levine, N., Matsukawa, A., & Ghasemzadeh, H. (2020, April). Improved knowledge distillation via teacher assistant. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 34, No. 04, pp. 5191-5198).
16. Micaelli, P., & Storkey, A. J. (2019). Zero-shot knowledge transfer via adversarial belief matching. Advances in Neural Information Processing Systems, 32.
17. Liu, P., Liu, W., Ma, H., Jiang, Z., & Seok, M. (2020, July). Ktan: knowledge transfer adversarial network. In 2020 International Joint Conference on Neural Networks (IJCNN) (pp. 1-7). IEEE.
18. Ye, J., Ji, Y., Wang, X., Gao, X., & Song, M. (2020). Data-free knowledge amalgamation via group-stack dual-gan. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 12516-12525).
19. Chung, I., Park, S., Kim, J., & Kwak, N. (2020, November). Feature-map-level online adversarial knowledge distillation. In International Conference on Machine Learning (pp. 2006-2015). PMLR.