A Review of Techniques for Enhancing Speech Recognition Systems

Aakash Anil Khatu, Caroline Anna Pasyanos, Deepesh Vejju, Rutvi Jiten Bheda

Abstract

We present a literature review of speech-to-text research in order to show recent improvements in the technology. We have discussed Listen, Attend, and Spell (LAS), a neural network that acquires the ability to translate spoken words into written characters. This model, in contrast to conventional DNN-HMM models, learns each component of a speech recognizer simultaneously. Two parts make up our system: a speller and a listener. The listener, which takes filter bank spectra as inputs, is a pyramidal recurrent network encoder. Characters are output by the speller, an attention-based recurrent network decoder. We also review SpecAugment, which aims to augment training data and improve accuracy in existing models. We also discuss how we can introduce a front-end support for existing ASR models to reduce noise as a preprocessor using existing Conformer Networks and increase efficiency and convergence of even large scale systems with a encoder decoder architecture. We also talk about Speaker invariant clustering (SPIN), which is a fine tuning method for STT models. SPIN disentangles the speaker information while preserving content representations.

Keywords: Speech to Text, Deep Neural Network(DNN), Hidden Markov Model(HMM), Automatic Speech Recognition(ASR), Conformer, Speaker Invariant Clustering

1. Introduction

Speech-to-text (STT), also referred to as automatic speech recognition or ASR, is a dynamic, evolving field of Davis et al. [5] introduced one of the first speech recognition systems in 1952, focusing on the development of an electronic circuit capable of identifying ten telephone-quality digits spoken at a normal speech rate. The circuit matched patterns by calculating the highest relative coefficient between each reference digit and the incoming data. Since then, STT has been used in a vast array of practical applications. STT is utilized by disabled students to enhance learning given by Matre and Cameron [10]. Another natural application is in multilingual communication provided in Ghai and Singh [8]. Much of current research efforts focus on increasing accuracy in STT models, several of which are reviewed in the following paper.

2. Listen Attend Spell Model

This model transcribes an audio sequence signal to a word sequence, one character at a time. It consists of a recurrent neural network RNN - listener and a decoder RNN called speller. It consists of a listener that converts low level speech into high level features and a speller which converts the above high level features into output expressions.

The major advantage of modeling characters as outputs gives the network to generate multiple spelling variants naturally. For example, for the phrase “triple a” the model produces both “triple a” and “aaa”.

2.1. Listen

A Bidirectional Long Short Term Memory RNN (BLSTM) is used in the Listen operation to handle problems with processing lengthy input voice signals. The reason to use an RNN is because it can take sequential data hence helping to keep track of past and present data. It is designed to remember and understand the sentences or conversations as mentioned in Chan et al. [2]. It can remember the sequence both forwards and backwards, meaning it can remember what was said before and what comes after, helping to understand the context. When the

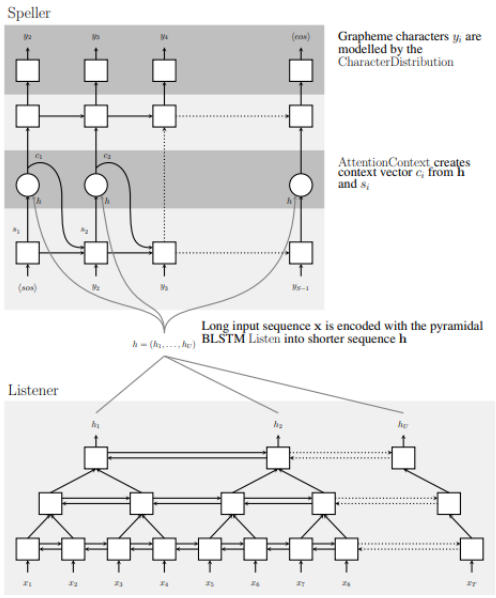

Figure 1: Listen, Attend and Spell (LAS) model: the listener is a pyramidal BLSTM encoding our input sequence x into high level features h and the speller is an attention-based decoder generating the y characters from h. Chan et al. [2]

voice is very long, the model struggles to handle it. Consider it like having a long story where we might mix up the parts or maybe forget it. To avoid this, we use a pyramid structure. This the variation that is made here so that we cut the long story in short chapters as mentioned in Chan et al. [2]. After processing each chapter, the key points are joined together before going to the next layer. Hence this keeps the story to be connected and understandable.

2.2. Attend and Spell

The Attend and Spell function is an attention-based LSTM transducer that computes a probability distribution over the next character at each output step, considering the characters seen previously. It is an attentive listener who picks up letter from a conversation and uses that information to predict what comes next. It looks at the letters that are already seen and guesses the next one. It is an attention mechanism which concentrates on specific parts of what is heard to figure out the next letter. It uses the information to predict what comes next.

2.3. Learning

The LAS function is hence trained together for an end to end speech recognition. However, because the model was not trained to be resilient to feeding in incorrect predictions at some time steps, the groundtruth is missing during inference, and the predictions may suffer as a result. We employ a technique that was suggested in Bengio et al. [1] to lessen this effect. During training, we occasionally take a sample from our previous character distribution and use that as the input for the next step predictions, rather than always feeding in the ground truth transcript.

The two functions are trained together in a way that they learn to recognize spoken words and then guess the next ones. Its like teaching a machine to both listen to a story and then retell it. Sometimes there might be issues such as not having the next correct words to refer to which may lead to mistakes as it has not learnt to guess without having the right answer to check against. In this case, instead of giving the next correct word, the system is made to guess based on its own previous guesses, helping to recover from its mistakes.

2.4. Decoding and Restoring

An algorithm similar to Sutskever et al. [14] that uses a basic left-to-right beam search is used for encoding. The goal here is to find the most likely sequence of characters and match the spoken words. To achieve this, the beam search method is used. Its like having a bunch of guesses(hypotheses) about what the words could be. You start with a simple guess and then keep adding characters to it and building a word letter by letter. At each step, you keep the most likely guesses and eliminate the less likely ones. A dictionary can be used to make sure the guesses are real words but it is found that the system usually guesses real words correctly without the help of a dictionary.

Introducing Listen, Attend, and Spell (LAS), an attention-driven neural network capable of directly translating auditory cues into character representations. The sequence to sequence framework serves as the foundation for LAS, and the encoder's pyramid structure minimizes the number of timesteps the decoder must handle. With two primary components, LAS is trained from start to finish. The input sequence is converted into a high level feature representation by the listener, the first component—a pyramidal acoustic RNN encoder. One character at a time, the second component, the speller, is an RNN decoder that handles the high level features and spells out the transcript.

3. SpecAugment

Previous research into improving STT software focused on designing better network architectures. Despite these efforts, the models developed overfit to the test data too easily and required a significantly large amount of training data. New research indicates that that effort may better be spent on augmenting input data as a means of creating additional training data. Building on previous research in this field, "SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition" by Park et al. [12] creates additional training data by manipulating the log mel spectrogram of input data as if it were an image. This method is computationally efficient to perform and does not require the collection of additional training data.

3.1. Data Augmentation

SpecAugment focused on three different types of log mel spectrogram deformations, each motivated by different aspects that could lead to inaccuracies in the STT model. The first deformation, time warping, contorts the log mel spectrogram image with the goal of training models to be more robust to time deformation in input data. Next, to make the model more robust to partial loss of frequency, the frequency masking deformation removes the data of some number of consecutive frequency channels. Finally, time masking removes the data of some number of consecutive timestamps.

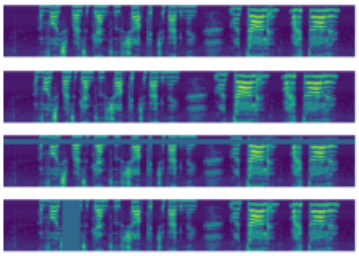

Figure 2: Augmentations applied to the base input, given at the top. From top to bottom, the figures depict the log mel spectrogram of the base input with no augmentation, time warp, frequency masking and time masking applied as provided in Park et al. [12]

This theoretically trains the model to be more robust in cases where small segments of speech are lost. The study also explored handcrafted policies where multiple frequency and time masks were applied to the same spectrograms, referred to as

LibriSpeech basic (LB), LibreSpeech double (LD), Switchboard mild (SM), and Switchboard strong (SS). Table 1 summarizes the augmentation parameters, and Figure 2 illustrates these policies.

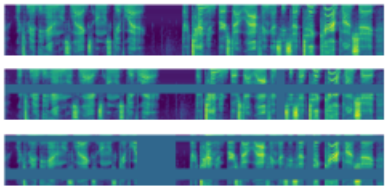

Figure 3: Augmentation policies applied to the base input. From top to bottom, the figures depict the log mel spectrogram of the base input with policies None, LB and LD applied. Park et al. [12]

Policy	W	F	m_F	T	p	m_T
None	0	0	-	0	-	-
LB	80	27	1	100	1.0	1
LD	80	27	2	100	1.0	2
SM	40	15	2	70	0.2	2
SS	40	27	2	70	0.2	2

Figure 4: Augmentation parameters for policies. mF and mT denote the number of frequency and time masks applied.Park et al. [12]

3.2. Experiment

As previously mentioned, SpecAugment builds upon the Listen, Attend, and Spell (LAS) class of end-to-end speech recognition models, as LAS networks are easily trained and have well-documented benchmarks. The paper reviews the learning rate schedules used in its research, finding them to be critical to further improving performance in speech recognition. It defines two learning schedules, B(asic) and D(ouble), as well as a L(ong) schedule used on the largest models. Park et al. [12] ran experiments using both LibriSpeech by Panayotov et al. [11] and Switchboard by Godfrey et al. [9] and with and without language models (LMs).

3.3. Results

With LibriSpeech 960h, the Listen, Attend, and Spell network and SpecAugment, 2.5-6.8% word error rate (WER) was achieved. With Switchboard 300h and SpecAugment (using both the SM and SS learning schedules) 6.8-14.4% WER was achieved. Both of these ranges outperform previous studies using the LibriSpeech and Switchboard. The study also concluded that training on augmented data caused the LAS network to underfit to the data where it previously overfitted. They argued that this was a benefit: under-fitting is addressed by making larger networks and training them for longer, which improves the network as a whole. In this instance, data augmentation successfully created a more accurate model without the need for additional data. This seems to suggest that data augmentation is a promising tool for creating more accurate models.

Method	**No LM**		**With LM**	
	clean	**other**	**clean**	**other**
HMM				
Panayotov et al., (2015) [20]			5.51	13.97
Povey et al., (2016) [30]			4.28	
Han et al., (2017) [31]			3.51	8.58
Yang et al. (2018) [32]			2.97	7.50
CTC/ASG				
Collobert et al., (2016) [33]	7.2			
Liptchinsky et al., (2017) [34]	6.7	20.8	4.8	14.5
Zhou et al., (2018) [35]			5.42	14.70
Zeghidour et al., (2018) [36]			3.44	11.24
Li et al., (2019) [37]	3.86	11.95	2.95	8.79
LAS				
Zeyer et al., (2018) [24]	4.87	15.39	3.82	12.76
Zeyer et al., (2018) [38]	4.70	15.20		
Irie et al., (2019) [25]	4.7	13.4	3.6	10.3
Sabour et al., (2019) [39]	4.5	13.3		
Our Work				
LAS	4.1	12.5	3.2	9.8
LAS + SpecAugment	**2.8**	**6.8**	**2.5**	**5.8**

Figure 5: LibriSpeech 960h WERs (%). Park et al. [12]

Method	**No LM**		**With LM**	
	SWBD	**CH**	**SWBD**	**CH**
HMM				
Veselý et al., (2013) [41]			12.9	24.5
Povey et al., (2016) [30]			9.6	19.3
Hadian et al., (2018) [42]			9.3	18.9
Zeyer et al., (2018) [24]			8.3	17.3
CTC				
Zweig et al., (2017) [43]	24.7	37.1	14.0	25.3
Audhkhasi et al., (2018) [44]	20.8	30.4		
Audhkhasi et al., (2018) [45]	14.6	23.6		
LAS				
Lu et al., (2016) [46]	26.8	48.2	25.8	46.0
Toshniwal et al., (2017) [47]	23.1	40.8		
Zeyer et al., (2018) [24]	13.1	26.1	11.8	25.7
Weng et al., (2018) [48]	12.2	23.3		
Zeyer et al., (2018) [38]	11.9	23.7	11.0	23.1
Our Work				
LAS	11.2	21.6	10.9	19.4
LAS + SpecAugment (SM)	**7.2**	14.6	**6.8**	14.1
LAS + SpecAugment (SS)	7.3	**14.4**	7.1	**14.0**

Figure 6: Switchboard 300h WERs (%). Park et al. [12]

4. Bringing the Noise

The existence of noise is a major obstacle for automated speech recognition (ASR) systems, which affects how well the systems function in practical settings. These applications cover a broad spectrum, such as call centers, real-time translation services, and voice assistants. In order to guarantee the resilience and dependability of ASR systems, scientists have created methods for purposefully adding synthetic noise to voice recordings. These techniques utilize a variety of strategies, including mimicking different noise levels by varying signal-to-noise ratios, imitating cross-talk and multilingual inclusions, and capturing background noises from sources like traffic or

music.

The work by Eickhoff et al. [6] suggests a strategy based on the usage of Conformers to address the crucial problem of noise in ASR. Conformers neural network architecture is known for its exceptional performance in ASR. This proposed architecture uses an encoder-decoder structure, with the encoder utilizing the Conformer architecture common to ASR systems and the decoder being specifically tailored to tasks like audio reconstruction and noise reduction.

4.1. Encoder

The main duty of the encoder is to extract latent features from the input data, thereby capturing important audio-related information such as language related context and phonetic context. A technique called "Parallel Weighted Sum" is introduced by the architecture to improve the extraction of these latent representations. This method uses several Conformers inside the encoder, and it processes each Conformer's output separately. These separate outputs then go via the layers of a neural network that are in charge of data transformation. Notably, the architecture computes a weighted sum rather than just combining these outputs. Several blocks with varying influences on the final representation are provided by this weighted sum. Because of its adaptability, the system can prioritize some features over others, which makes it especially useful for resolving issues with noise in ASR.

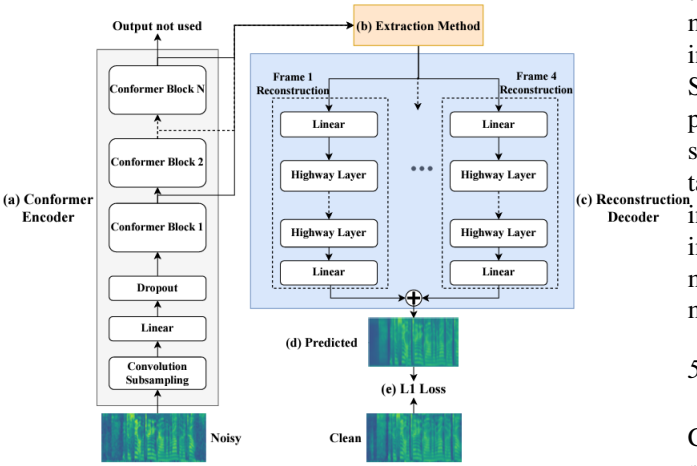

Figure 7: Encoder-Decoder architecture to clean a noisy signal and produce a clear signal as proposed by Eickhoff et al. [6]

4.2. Decoder

The primary goal of the architecture's decoder component is to enhance the denoised spectrograms, which are graphical depictions of sound signals. In order to accomplish this, the architecture makes use of Highway Networks, a particular kind of neural network that is well-known for its capacity to successfully recreate spectrograms. Similar to Long Short-Term Memory (LSTM) networks, Highway Networks feature gated connections. By controlling the information flow through the network, these gate mechanisms attenuate noise and irrelevant information and allow for the collection of critical features. Given that the Conformer preprocessing block reduces the temporal dimension of the data by a factor of four, four distinct Highway Networks are trained. Each of these networks specializes in handling different aspects of the processing of the denoised data, and they operate independently. The outputs generated by these four networks are then concatenated along the temporal axis, resulting in a comprehensive denoised representation that effectively captures various aspects of the speech signal over time.

In summary, this novel architecture uses the capability of Conformers within an encoder-decoder framework to overcome the noise challenge in ASR. The usage of Highway Networks in the decoder optimizes the denoising and reconstruction process, while the introduction of the "Parallel Weighted Sum" approach allows flexibility and noise reduction in latent representation extraction. When these methods are combined, the audio signal is robustly denoised, which enhances the accuracy and performance of ASR systems in noisy settings and real-world applications. This model could be used as a fronted-end architecture for many encoder-decoder based architecture to provide better results with a faster convergence.

5. SPIN

In recent research, there are a lot more ways to improve the accuracy of STT models which Involve modifying the architecture of the models itself instead of training the models on more robust datasets, once such method proposed is Speaker invariant clustering (SPIN), which is a fine tuning method for STT models. SPIN disentangles the speaker information while preserving content representations. These disentangled representations of the spoken content improve various downstream tasks such as ASR. SPIN only required 45 minutes of fine tuning on a single GPU as given by Chang et al. [3] which is a huge improvement when compared to ContentVec which is a similar method but takes 19 hours on 36 GPUs on top of the pre-trained models as shown by Choi et al. [4].

5.1. Methodology

Speaker Perturbation. SPIN uses an algorithm proposed by Choi et al. [4] as ContentVec Qian et al. [13] . The Algorithm randomly and uniformly scales the formant frequencies and F0 as proposed by Eide and Gish [7] resulting in the alteration of speaker's voice information with little content loss.

The process involves two main components: Speaker Invariant Clustering and Speaker Invariant Swapped predictions, the complete architecture is as shown in figure 8. First a speaker augmented speech is created by adding speaker perturbations, Then the augmented and original speech pair views are linearly projected and L2 normalized into representations.

The First component Speaker Invariant Clustering, starts off by calculating a probability distribution for each frame of speech to represent the similarity between the input and the learnable codewords in the codebook. The goal is to find consistent content across different speakers which means that the distribution of codewords over the codebook should Ideally be similar hence the goal is to minimize the cross entropy.

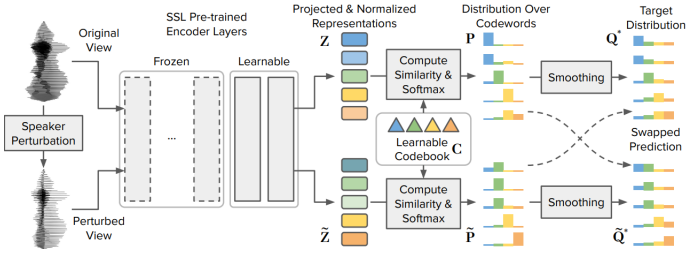

Figure 8: The Spin architecture. A new view is generated with a simple speaker perturbation. A pre-trained speech SSL model extracts representations from both utterances (Z/ Z˜). Representations are projected, normalised, and quantized with a learnable codebook into probability distributions (P/ P˜). The distributions are smoothed to enforce full codebook usage (Q / Q˜). Finally, each frame's distribution is used to predict the target distribution produced by the other view (P → Q˜ and P˜ → Q). from Chang et al. [3]

To avoid all representations being clustered into a single codeword, the target distribution is smoothed. This encourages the use of all codewords in the codebook more evenly.

The second component is speaker-invariant swapped prediction, which swaps the roles of augmented and original speech pairs. The goal here is to encourage the model to produce similar codeword representation at the same position between two different views.

The SPIN is applied to pre-trained STT models as using SPIN to train models from scratch would only learn the positional information in the speech representations.Another benefit of SPIN is that it does not require random masking, thus utilizing all frames for network updates.

5.2. Experiments

Spin is trained on the LibriSpeech 100 hour subset and the paper reports that using more data does not improve. The paper then compares accuracy in tasks like phoneme recognition, automatic speech recognition, keyword spotting, query by example, intent classification and slot filling, testing Pretrained models HuBERT and WavLM without SPIN, ContentVec, wav2vec 2.0, daata2vec and HuBERT and WavLM models with SPIN applied. These models are evaluated first in the speech processing universal performance benchmark (SUPERB) and then on the zero resource speech benchmark for comparing the Acoustic Unit Discovery and Discreet Unit Quality. Speaker Invariance : when tested on the SUPERB speaker identification test, SPIN reduces identification accuracy to 10% in the last layer which signifies that SPIN is able to completely remove the speaker's speech characteristics from the spoken content.

5.3. Summary

The method proposed by the paper, SPIN, is able to improve the accuracy of pretrained models on various content related tasks while adding only a relatively small amount of computational cost in the pre-training. The paper mentions that although currently SPIN is only implemented on HuBERT and WavLM, it can also be applied to enhance other SSL models.

6. Conclusion

This paper has explored recent advancements in speech-to-text (STT) models, with a focus on noise robustness and speaker invariance. We've discussed the Listen, Attend and Spell model, an end-to-end neural network for direct speech-to-text transcription, and two methods for enhancing its performance by modifying the training data. SpecAugment builds upon the Listen, Attend and Spell model and presents a data augmentation method that improved WER across several language models. Bring the Noise augmented input data by removing noise before passing it through the model, also resulting in higher accuracy. We also discussed Speaker Invariant Clustering(SPIN), a method for speaker disentanglement. The results demonstrate an increase in performance and accuracy on various speech recognition tasks. We hope that this paper will provide a useful overview of recent improvements in STT technology, and serves as a reference for researchers and practitioners who are interested in developing and applying STT models.

References

[1] Bengio, S., Vinyals, O., Jaitly, N., Shazeer, N., 2015. Scheduled sampling for sequence prediction with recurrent neural networks. Advances in neural information processing systems 28.

[2] Chan, W., Jaitly, N., Le, Q.V., Vinyals, O., 2015. Listen, attend and spell. arXiv preprint arXiv:1508.01211 .

[3] Chang, H.J., Liu, A.H., Glass, J., 2023. Self-supervised fine-tuning for improved content representations by speaker-invariant clustering. arXiv preprint arXiv:2305.11072 .

[4] Choi, H.S., Lee, J., Kim, W., Lee, J., Heo, H., Lee, K., 2021. Neural analysis and synthesis: Reconstructing speech from self-supervised representations. Advances in Neural Information Processing Systems 34, 16251–16265.

[5] Davis, K.H., Biddulph, R., Balashek, S., 1952. Automatic recognition of spoken digits. The Journal of the Acoustical Society of America 24, 637–642.

[6] Eickhoff, P., Möller, M., Rosin, T.P., Twiefel, J., Wermter, S., 2023. Bring the noise: Introducing noise robustness to pretrained automatic speech recognition, in: International Conference on Artificial Neural Networks, Springer. pp. 376–388.

[7] Eide, E., Gish, H., 1996. A parametric approach to vocal tract length normalization, in: 1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings, IEEE. pp. 346–348.

[8] Ghai, W., Singh, N., 2012. Literature review on automatic speech recognition. International Journal of Computer Applications 41.

[9] Godfrey, J.J., Holliman, E.C., McDaniel, J., 1992. Switchboard: Telephone speech corpus for research and development, in: Acoustics, speech, and signal processing, ieee international conference on, IEEE Computer Society. pp. 517–520.

[10] Matre, M.E., Cameron, D.L., 2022. A scoping review on the use of speech-to-text technology for adolescents with learning difficulties in secondary education. Disability and Rehabilitation: Assistive Technology , 1–14.

[11] Panayotov, V., Chen, G., Povey, D., Khudanpur, S., 2015. Librispeech: an asr corpus based on public domain audio books, in: 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP), IEEE. pp. 5206–5210.

[12] 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 .

[13] Qian, K., Zhang, Y., Gao, H., Ni, J., Lai, C.I., Cox, D., Hasegawa-Johnson, M., Chang, S., 2022. Contentvec: An improved self-supervised

speech representation by disentangling speakers, in: International Conference on Machine Learning, PMLR. pp. 18003–18017.
[14] Sutskever, I., Vinyals, O., Le, Q.V., 2014. Sequence to sequence learning with neural networks. Advances in neural information processing systems 27.