Exploring the impact of noise in low-resource ASR for Tamil

Vigneshwar Lakshminarayanan
Walmart Global Tech
vicky18.kb@gmail.com

Emily Prud'hommeaux
Boston College
prudhome@bc.edu

Abstract

The use of deep learning algorithms has resulted in significant progress in automatic speech recognition (ASR). Robust high-accuracy deep neural ASR models typically require thousands or tens of thousands of hours of speech data, but even the strongest models can fail under noisy conditions. Unsurprisingly, the impact of noise on accuracy is more dramatic in low-resource settings. In this paper, we investigate the impact of noise on ASR in a low-resource setting. We explore novel methods for developing noise-robust ASR models using a small dataset for Tamil, a widely-spoken but under-resourced Dravidian language. We add various noises to the audio data to determine the impact of different kinds of noise (e.g., punctuated vs. continuous, mechanical vs natural). We also explore whether different data augmentation methods are better suited to handling different types of noise. Our results show that all noises, regardless of the type, had an impact on ASR performance, and that upgrading the architecture alone could not fully mitigate the impact of noise. In our experiments, SpecAugment, a common data augmentation method for end-to-end neural ASR, was not as helpful as raw data augmentation, in which noise is explicitly added to the training data. Raw data augmentation enhances ASR performance on both clean data and noise-mixed data.

1 Introduction

Automatic Speech Recognition (ASR) technology is widely used in many modern applications for high-resource languages, such as dictation and personal assistants like Amazon Alexa and Apple's Siri (Yoshioka et al., 2012). The success of ASR for these applications is due largely to the emergence of deep learning architectures, improvements in computing hardware, and the large amounts of data available for languages like English and Mandarin (Ruan et al., 2018). The performance of even high-accuracy ASR models, however, remains fragile in the presence of external noise. ASR accuracy degrades even further in low-resource settings.

The motivation for this paper is to research the impact of several types of noise (e.g., continuous vs. punctuated, mechanical vs. natural) on ASR performance in a low-resource setting for the Dravidian language, Tamil. We explore a range of ASR architectures, including traditional GMM-HMM (Reynolds, 2009), SGMM, (Povey et al., 2010), and a hybrid DNN model (Veselý et al., 2013). Additionally, we evaluate whether different data augmentation approaches, such as raw audio augmentation and spectral augmentation (SpecAugment) (Park et al., 2019), are particularly well suited to different types of noise. We explore these questions using a low-resource dataset for Tamil provided by SpeechOcean.com and Microsoft for a low-resource ASR challenge in Interspeech 2018 (Srivastava et al., 2018). We observe that all types of noises, regardless of the acoustic model architecture, degrade ASR performance. Although not entirely surprising, we also find that raw audio augmentation outperforms the popular SpecAugment (Park et al., 2019) data augmentation method on clean data as well as noise-mixed data.

2 Previous Work

There is some prior work on using noise-mixed data to make ASR more robust to noise and other external conditions, but most of this work focuses on high-resource languages. Pervaiz et al. (2020) provided a comparative study on various acoustic and deep learning models, creating robust models in a noisy environment. The models were trained on noise-augmented training data and tested on both clean and noisy data. Hu et al. (2021) proposed a noise-robust speech recognition system called Interactive Feature Fusion Fetwork (IFF-Net) to learn the missing latent information by combin-

Proceedings of the Fourth Workshop on Speech, Vision, and Language Technologies for Dravidian Languages, pages 30–34

ing the enhanced features and the original noisy features into a fused representation. This system achieved better results on the RATS Channel-A corpus. Shrawankar and Thakare (2013) investigated challenges due to changes in environmental conditions and speaker characteristics and proposed a method to increase the robustness of the ASR systems using speech enhancement techniques like spectral normalization and spectral subtraction. Giurgiu and Kabir (2011) explained how energy normalization and speech re-synthesis can improve the performance of ASR systems by recognizing speech signals in high-noisy conditions (negative SNR). Kinoshita et al. (2020) investigated whether the usage of single-channel time-domain neural networks can help in the reduction of noise and thereby improve the performance. Gupta et al. (2016) proposed a Back-propagation Artificial Neural Network with feature compression using MFCCs yielding improved performance with low signal-to-noise ratios.

3 Data

The data used for the approach is a low-resource Tamil language dataset, provided by SpeechOcean.com and Microsoft for a low-resource ASR challenge for Indian languages at Interspeech 2018 (Srivastava et al., 2018). The dataset consists of read speech and conversations that have been split into utterances and transcribed. The dataset contains a total of 50 hours of recorded speech data in a clean noise-free environment. The dataset is partitioned into 40 hours of speech data for training, 5 hours for testing, and 5 hours for development. We do not use the dev set in our work. All the audio files are 16-bit mono audio sampled at 16kHz. A total of 1900 speakers are included in the dataset. Utterances range in length from 3000ms to 10000ms. There are a total of 42,212 unique utterances in the dataset.

4 Methodology

We explore three different ASR architectures available with the Kaldi speech recognition toolkit (Povey et al., 2011). We note that none of these can be considered state-of-the-art. However, prior work has shown that the Kaldi hybrid DNN is competitive with fine-tuning from multilingual end-to-end models using wav2vec XLSR (Baevski et al., 2020) and Whisper (Radford et al., 2022) in low-resource settings (Jimerson et al., 2023).

GMM-HMM Acoustic Model The Gaussian Mixture Model (GMM) is used to estimate the probability density function used in statistical classification systems (Reynolds, 2009). GMMs are commonly used in statistical ASR to estimate likelihoods of phones (speech sounds) given their acoustic features (typically MFCCs, Mel-frequency cepstral coefficients). Combined with Hidden Markoc Models (HMMs), GMMs are used to estimate the density and maximize the likelihood of the distribution of the speech sounds.

Subspace GMM Acoustic Model A subspace GMM is a type of acoustic model where all the phoneme states use a common Gaussian Mixture Model structure (Povey et al., 2010). The SGMM model is trained by clustering the Gaussians from the GMM-HMM model using the Universal Background Model (UBM). The UBM model is a speaker-independent high order GMM model(Povey et al., 2008). The SGMM models are trained using the UBM model with the state probability distribution functions identical. The final step of the training process is to use the EM algorithm to train the SGMM model using the alignments from the GMM-HMM and also from the SGMM model as well (Povey et al., 2010).

Hybrid Deep Neural Network Acoustic Model As noted above, more recent research has turned from statistical ASR to Deep Neural Networks (DNN). Here we use a relatively simple fully-connected feed-forward DNN using "Karel's implementation" (Veselý et al., 2013) within the Kaldi toolkit. (Povey et al., 2011). The DNN model consists of 6 hidden layers where each hidden layer has 2048 nodes (Cosi, 2015). The DNN model is trained using the features extracted in the GMM-HMM acoustic model described above, yielding a hybrid, rather than end-to-end, architecture.

4.1 Data Augmentation

Data augmentation is the process of including additional data in the ASR training data set with the goal of improving performance by increasing and diversifying the training data. Usually, the added data is created synthetically – either by modifying the existing training data in some way or by generating new data through speech synthesis. Here we focus on modifying the existing data through raw audio augmentation and through spectral augmentation.

Model	Clean WER	Continuous Natural		Punctuated Natural		Continuous Mechanical		Punctuated Mechanical	
		Party	*Restaurant*	*Dog*	*Cat*	*Tap*	*Dishes*	*Truck Horn*	*Door Slam*
GMM-HMM	44.66	66.23	53.20	52.20	48.0	61.2	55.1	50.52	50.21
SGMM	36.15	66.05	47.75	45.19	41.26	54.4	48.18	45.24	44.94
DNN	32.58	56.88	41.18	39.91	35.91	47.18	41.54	39.44	39.12

Table 1: Results for test data mixed with the eight noises using ASR models trained on unaugmented training data.

Augmentation Noise	Model	Clean WER	Continuous Natural		Punctuated Natural		Continuous Mechanical		Punctuated Mechanical	
			Party	*Restaurant*	*Dog*	*Cat*	*Tap*	*Dishes*	*Horn*	*Door Slam*
None	DNN	32.58	56.88	41.18	39.81	35.91	47.18	41.54	39.44	39.12
	SGMM	36.15	66.05	47.75	45.19	41.26	54.4	48.18	45.23	44.94
Tap and Dishes	DNN	31.88	52.84	37.81	38.83	35.09	38.03	36.21	38.16	38.27
	SGMM	36.16	63.42	45.87	44.34	41.02	48.81	44.23	44.38	42.87
Horn and Door	**DNN**	31.47	55.79	39.91	38.83	33.23	35.88	39.17	36.06	35.88
	SGMM	35.19	63.47	46.15	44.13	40.9	52.88	46.67	40.97	40.2
Party and Restaurant	**DNN**	31.85	46.16	35.18	37.35	33.87	38.41	36.91	35.26	38.41
	SGMM	35.84	57.27	41.7	43.1	39.35	50.71	44.48	41.18	44.03
Dog and Cat	**DNN**	31.72	54.85	38.97	34.26	32.07	38.11	39.83	38.24	38.11
	SGMM	35.49	64.0	45.73	38.25	36.59	53.18	46.8	43.89	42.25
SpecAug	**DNN**	35.75	58.86	45.03	46.27	39.53	50.63	43.97	44.81	42.74
	SGMM	43.06	68.56	54.27	51.68	47.11	64.62	54.52	52.32	51.31

Table 2: Results for test data mixed with the eight noises (columns) using ASR models trained on data augmented with the eight different noises (rows), using the two strongest architectures, DNN and SGMM.

4.1.1 Spectral Augmentation

In Spectral Augmentation (SpecAugment) (Park et al., 2019) random sections of the spectral representation of a speech sample are set to zero. It is performed using the log Mel spectrogram of the input speech data. SpecAugment is widely used because of its simplicity and effectiveness within neural end-to-end ASR frameworks. It does not require knowledge of the phonetic content of the speech signal, and because it is applied to a spectral representation and simply involves reducing to zero, it is computationally inexpensive. We note that SpecAugment was designed for end-to-end neural ASR rather than the statistical and hybrid architectures we explore in this paper, so its performance in our models may not be ideal.

4.1.2 Raw Augmentation

Raw augmentation involves directly modifying the raw audio signal. Here, in order to investigate the differential impact of noise, we create a new version of the existing training data by mixing various noise samples (e.g., faucet running, cocktail party chatter, dog barking) into the existing audio. We refer to the resulting datasets as the noise-mixed augmented training sets. A small set of noises, again 16-bit mono sound files samples at 16kHz, were selected from an existing noise dataset. We experimented with two dimensions of noise: mechanical vs. natural, and punctuated vs. continuous. The categorization across these dimensions of the eight sounds used is shown in Table 3. Each sound was decreased by 20dB and then superimposed on the existing data.

	Continuous	**Punctuated**
Mechanical	Tap running Washing dishes	Truck horn Door slam
Natural	Restaurant Party chatter	Dog bark Cat meow

Table 3: The eight sounds used in the project, categorized in the two dimensions: continuous vs. punctuated, and mechanical vs. natural.

For each speech file (utterance) in the test set, there are 9 versions: one clean (i.e., the original data), and one with each of the eight specified noises. The unaugmented training set consists solely of the original training data. There are 4 raw audio augmented (noise-mixed) training sets, one for each of the four categories shown in Table 3. Each of these four training sets consists of one clean copy of each training utterance, one copy of that utterance with one of the relevant noises in that category (e.g., one mechanical continuous noise, tap running) superimposed, and one copy with the other relevant noise in that category (e.g., the other mechanical continuous noise, dish washing) superimposed. Finally, the sixth training set, for exploring SpecAugment, consists of a clean copy of each utterance and a copy that has been passed through SpecAugment with the parameters $F = 30, T = 40, m_T = m_F = 2$.

4.2 Evaluation Metric

We use the standard ASR evaluation metric, Word Error Rate (WER), to assess the performance of the ASR models under different training and testing conditions. WER is calculated as the number of insertions, deletions, and substitutions in the hypothesized ASR output relative to a reference transcript divided by the number of words in the reference transcript. A lower WER indicates higher ASR accuracy.

5 Results

The baseline model was trained using the 40 hours of unaugmented training data and tested using the five hours of test data with no modification or augmentation of the test or train data. In the **Clean WER** column in Table 4, we see the baseline performance for the top-performing architectures discussed above: GMM-HMM, SGMM, and DNN, all trained with the Kaldi ASR toolkit (Povey et al., 2011). The remaining columns of that table show the performance of these models, trained on unaugmented data, on test data with the 8 noises inserted.

We see that, not unsurprisingly, the DNN architecture produces the best (lowest) WER, with the GMM-HMM acoustic model having the weakest performance. The results on test data mixed with the eight noises show that adding noises degrades ASR performance regardless of architecture and sound type, sometimes very dramatically. The cat meowing has the smallest impact of the 8 noises, while party chatter has the largest.

Table 2 shows that the impact of noise can be reduced by augmenting the training data via noise mixing, as desribed above. Every raw audio augmentation approach reduces WER, even if the noise type differs on one or both dimensions. It appears that adding any noise at all to the training data will yield improvements in WER, regardless of the architecture used. Raw audio data augmentation consistently outperforms the widely used SpecAugment augmentation technique, but we note that SpecAugment is intended to be used with neural end-to-end ASR architectures.

6 Conclusion

The goal of this research is to investigate the impact of noise on Automatic Speech Recognition models using a low-resource Tamil language dataset. We investigated the impact of noise by mixing several kinds of noises into the testing data. We evaluated the performance on the baseline model trained exclusively on the original unaugmented data. We discovered that all noises, regardless of whether they were mechanical or natural, continuous or punctuated, degraded ASR performance, and that upgrading the architecture alone was unable to fully mitigate the impact of noise.

To reduce the impact of noise in ASR models, we augmented the training data by superimposing noises from each of the four categories onto the training data. We also employed the popular spectral augmentation technique, SpecAugment to create another augmented training dataset. We discovered that raw data augmentation improves WER, regardless of the combination of noises in the training and test sets. We also found that targeted raw augmentation improves ASR performance: adding noise to the training data that shares one or more characteristics with the noise in the test data yields larger improvements.

Our methods outperform SpecAugment, although we recognize that SpecAugment is designed for end-to-end neural ASR architectures rather than statistical or hybrid architectures like those explored here. In our future work, we plan to investigate these augmentation methods with an end-to-end architectures, with a particular focus on approaches that involve fine-tuning multilingual models to low resource datasets, including wav2vec XLSR (Baevski et al., 2020) and Whisper (Radford et al., 2022).

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460.

Piero Cosi. 2015. A kaldi-dnn-based asr system for italian. pages 1–5.

Mircea Giurgiu and Ahsanul Kabir. 2011. Improving automatic speech recognition in noise by energy normalization and signal resynthesis. In *2011 IEEE 7th International Conference on Intelligent Computer Communication and Processing*, pages 311–314.

Santosh Gupta, Kishor M. Bhurchandi, and Avinash G. Keskar. 2016. An efficient noise-robust automatic speech recognition system using artificial neural networks. In *2016 International Conference on Communication and Signal Processing (ICCSP)*, pages 1873–1877.

Yuchen Hu, Nana Hou, Chen Chen, and Eng Siong Chng. 2021. Interactive feature fusion for end-to-end noise-robust speech recognition. *arXiv preprint arXiv:2110.05267*.

Robert Jimerson, Zoey Liu, and Emily Prud'hommeaux. 2023. An (unhelpful) guide to selecting the best asr architecture for your under-resourced language. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1008–1016.

Keisuke Kinoshita, Tsubasa Ochiai, Marc Delcroix, and Tomohiro Nakatani. 2020. Improving noise robust automatic speech recognition with single-channel time-domain enhancement network. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7009–7013.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin Dogus Cubuk, and Quoc V. Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. In *INTERSPEECH*.

Ayesha Pervaiz, Fawad Hussain, Humayun Israr, Muhammad Ali Tahir, Fawad Riasat Raja, Naveed Khan Baloch, Farruh Ishmanov, and Yousaf Bin Zikria. 2020. Incorporating noise robustness in speech command recognition by noise augmentation of training data. *Sensors (Basel, Switzerland)*, 20.

Daniel Povey, Lukśš Burget, Mohit Agarwal, Pinar Akyazi, Kai Feng, Arnab Ghoshal, Ondřej Glembek, Nagendra Kumar Goel, Martin Karafiát, Ariya Rastrow, Richard C. Rose, Petr Schwarz, and Samuel Thomas. 2010. Subspace gaussian mixture models for speech recognition. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4330–4333.

Daniel Povey, Stephen M. Chu, and Balakrishnan Varadarajan. 2008. Universal background model based speech recognition. In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 4561–4564.

Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al. 2011. The kaldi speech recognition toolkit. In *IEEE 2011 workshop on automatic speech recognition and understanding*, CONF. IEEE Signal Processing Society.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

Douglas A. Reynolds. 2009. Gaussian mixture models. In *Encyclopedia of Biometrics*.

Sherry Ruan, Jacob O Wobbrock, Kenny Liou, Andrew Ng, and James A Landay. 2018. Comparing speech and keyboard text entry for short messages in two languages on touchscreen phones. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 1(4):1–23.

Urmila Shrawankar and Vilas M. Thakare. 2013. Adverse conditions and ASR techniques for robust speech user interface. *CoRR*, abs/1303.5515.

Brij Mohan Lal Srivastava, Sunayana Sitaram, Rupesh Kumar Mehta, Krishna Doss Mohan, Pallavi Matani, Sandeepkumar Satpal, Kalika Bali, Radhakrishnan Srikanth, and Niranjan Nayak. 2018. Interspeech 2018 Low Resource Automatic Speech Recognition Challenge for Indian Languages. *Proc. 6th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2018)*, pages 11–14.

Karel Veselý, A. Ghoshal, Lukas Burget, and Daniel Povey. 2013. Sequence-discriminative training of deep neural networks. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, pages 2345–2349.

Takuya Yoshioka, Armin Sehr, Marc Delcroix, Keisuke Kinoshita, Roland Maas, Tomohiro Nakatani, and Walter Kellermann. 2012. Making machines understand us in reverberant rooms: Robustness against reverberation for automatic speech recognition. *IEEE Signal Processing Magazine*, 29(6):114–126.