BANDPASS NOISE GENERATION AND AUGMENTATION FOR UNIFIED ASR

Kshitiz Kumar, Bo Ren, Yifan Gong, Jian Wu

Microsoft Corporation, Redmond, WA
{kshitiz.kumar, boren, ygong, jianwu }@microsoft.com

ABSTRACT

Data Simulation is a crucial technique for robust automatic speech recognition (ASR) systems. We develop this work in the scope of data augmentation and improve robustness by generating new bandpass noise resources from an existing noise corpus. We design numerous bandpass filters with varying center frequencies and filter bandwidths, and obtain corresponding bandpass noise samples. We augment our baseline data simulation with bandpass noises to ingest additional robustness and generalization to generic and unknown acoustic scenarios. This work targets ASR robustness to individual subband noises, and improves robustness to unseen real-world noise that can be approximated as a factorial combination of subband noises. We demonstrate our work for a large scale unified ASR task. We obtained 7% word error rate relative reduction (WERR) across unseen acoustic conditions and 11% WERR for kids speech. We also demonstrate generalization to new ASR applications.

Index Terms— LSTM, Data Simulation, Digital Assistant, Robust ASR, Kids speech

1. INTRODUCTION

We are witnessing many on-device and on-cloud speech applications that deliver strong ASR performance across a range of applications. Deep learning techniques have led to speech-enabled smart assistants in Cortana, Alexa, Google Home and Siri. ASR users have greater expectations from speech products and expect it to work equally well across numerous acoustic environments, including noisy conditions, far-field, non-native speech, kids, natural conversation, and side-speech etc. Current ASR systems are trained on large scale training tasks, still above expectations strain the robustness requirements of our speech products, and encourage us to innovate and inculcate new robustness in ASR.

Over the past years, speech researchers have developed a variety of algorithms and architectures to train robust ASR models. The deep long-short term memory (LSTM) models in [1, 2, 3, 4] have demonstrated improvements over an earlier application of deep learning in DNN models [5, 6, 7, 8]. Recent advances in deep learning also include end-to-end systems in [9].

ASR systems also benefit from model or speaker adaptation [10, 11, 12, 13] that personalize models towards a specific acoustic scenario or speaker. Beamforming and stream combination techniques in [14, 15] too improve ASR robustness. Furthermore, acoustic model combination techniques promise a single unified model to broaden the unified ASR reach to many seen as well as unseen acoustic conditions. Some examples include hypothesis combination with ROVER [16], using confidence scores [17], and classifier-based system combinations in [18]. There has also been significant development in inventing robust ASR features. [19] provides a recent overview of robust techniques.

Data simulation too is crucial for training robust models [20, 21, 22, 23, 24, 25, 26], and has been demonstrated to be effective. In aforementioned work, data simulation and augmentation have been applied to low-resource languages, as well as, new application scenarios with minimal application-dependent training data. Data simulation provides a way to reuse training data from potentially different scenarios by appropriately simulating it for particular target application. The techniques are effective for small, as well as, large scale ASR training tasks.

Our proposed work builds in the scope of new data simulation techniques for robust ASR. Our baseline system already incorporates the benefits of data augmentation from real-world noise examples. This work develops on that and motivates robustness to individual frequency bands. Towards that, we design and generate a large variety of bandpass filters across center frequencies and filter bandwidths. We apply the bandpass filters to an existing noise corpus and generate multiple bandpass noise samples. Subsequently we apply subband noises for data augmentation on top of our baseline data augmentation system. We hypothesize that ASR training with subband noise samples builds robustness to unseen real-world acoustic conditions, as they can be approximated by a combinations of individual subband noises. This can directly improve ASR systems to new acoustic conditions.

Next we provide a motivation for robust ASR in Sec. 2. We develop our bandpass noise simulation work in Sec. 3. We set up experiments to evaluate our work in Sec. 4 and present corresponding results. We also discuss our findings in Sec. 5, and conclude this study in Sec. 6.

2. MOTIVATION FOR UNIFIED ASR

In this section, we motivate the robustness and generalization requirements of ASR systems. We aspire to build a unified ASR service for diverse usage. As ASR service providers, we may not have prior information about individual scenarios, so our solutions should extend to all possible ASR applications. We highlight the ASR robustness expectations for, (a) ASR consumers that may include adults or kids; natural conversation or command/control; non-native speakers and speech accents, (b) acoustic environments including background noise, distant speech devices in far-field environments, (c) speech-enabled devices with device-specific signal acquisition and processing, including narrowband or wideband audio, (d) communication networks including data encoding and quantization techniques. It's understandable that despite all the advances in large scale deep learning solutions, robustness is still a challenge for unified ASR.

An effective approach for robustness to a particular acoustic scenario is to collect training data from that scenario, and add to ASR training. The deep learning models are surprisingly good at memorizing diverse training data. Although useful, above approach is costly or infeasible in many applications. Any data collection requires very careful design to account for the diversity in speakers and acoustic environments. Furthermore, the target devices may be in parallel development, making the collected data outdated at a later point. We have also seen differences in above "engineered" data collection and that from live product usage.

Data simulation techniques provide effective ASR robustness, and has been demonstrated across applications [24]. These techniques leverage existing data resources and re-purpose data for target application domains. Some specific examples include far-field data simulation from near-field speech [23], or noisy speech simulation from clean speech corpus. Other examples include VTLP [21] and an application to low-resource languages in [20]. Next in Sec. 3, we develop our work for substantial ASR robustness on top of an already robust baseline.

3. BANDPASS NOISE SIMULATION FOR ASR

We have noted that data simulation is a critical technique and widely used in ASR community. It's standard to train ASR with data augmentation from noise and room impulse responses (RIR) for ASR robustness. That makes our models considerably more robust than training without data augmentation. Despite those gains, we still find applications with greater robustness requirements, *e.g.*, data from new far-field devices not seen in training, or an application to unseen acoustic conditions. Furthermore, a stronger data augmentation technique also improves the core ASR accuracy for scenarios with limited training data, some examples include accented and kids speech.

This work focuses on above aspects and motivates bandpass noise simulation to make ASR specifically robust to noise in individual frequency bands. This leads to greater richness in noise simulation; we also expect the ASR model to develop robustness to individual subband noises, as well as, the factorial combinations of subband noises. Alternatively we hypothesize that the trained model can generalize well to unseen noises that can be represented as a combination of numerous subband noise samples. This promises to substantially increase the ASR robustness to new acoustic conditions.

We describe our data simulation work with respect to Fig. 1. There "baseline" training denotes ASR training without data augmentation. We refer to "Standard Sim." as ASR training by augmenting ASR training corpus with standard data augmentation from an available noise corpus. For bandpass simulation we feed the noise corpus through bandpass filters at different center frequencies and bandwidths. "Bandpass Sim." training in Fig. 1 includes: (a) original ASR corpus, (b) data augmentation from a noise corpus, and (c) data augmentation from bandpass noise samples generated from a noise corpus.

3.1. Bandpass Filter Design

The bandpass filters for data simulation can be designed in numerous ways. For current work we chose 2-pole Butterworth bandpass filters. We note a particular bandpass filter with 3-dB band-width as B-Hz at center frequency C-Hz. For this work we designed the set:

$$B \in \{200, 300, 400\} \tag{1}$$
$$C \in \{200, 300, 400, ..., 7500\} \tag{2}$$

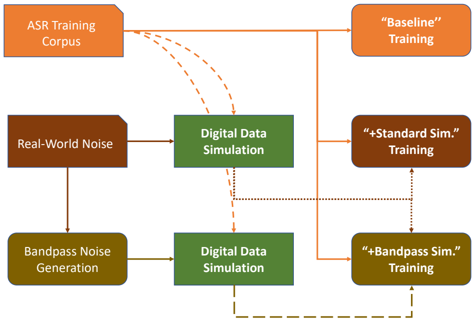

Fig. 1. *ASR training for Baseline, Standard Simulation and Bandpass Simulation techniques. Note that Bandpass Simulation includes: (a) original ASR training data, (b) data augmentation from a noise corpus, and, (c) augmentation from bandpass noise samples.*

We select a noise sample from our noise corpus, obtain 8-16 randomly selected pairs (B, C) from above set, and generate corresponding bandpass noise samples. We iterate above procedure to all data in the noise corpus. Note that the selected set (B, C) will differ across individual noises in the noise corpus, leading to significant diversity in the generated bandpass samples. Above procedure has an effect of oversampling the noise corpus by 8-16 times, and starting from 1000 noise samples we produce upto 16000 unique bandpass samples. We believe above set of B and C provides adequate coverage for ASR robustness. We also do not anticipate a need to increase our current oversampling rate. This work focuses on smaller subbands so an upper limit of B as 400-Hz for 3-dB band-width provides a reasonable choice.

We also demonstrate a particular noise sample and corresponding bandpass noise samples from bandpass filters at different center frequencies (C) in Fig. 2. By incorporating a rich and diverse set of bandpass noise samples in data simulation, we expect our work to inculcate new ASR robustness to subband noises, as well as, real-world noise conditions.

4. EXPERIMENTS AND RESULTS

We conduct our experiments on a large vocabulary speech recognition task. We build a standard 6-layers unidirectional LSTM model with cross-entropy (CE) [5] criterion from a large data corpus from Microsoft speech services. This is anonymized data with personally identifiable information removed. With an objective to evaluate the robustness of new scenarios, we design our training set to include a mix of near-field as well as Xbox (far-field A), with a total of approximately 2500 hrs speech. This data includes a mix of command and query, dictation, and natural conversation. Our training noise corpus consists of over 1000 real-world noise samples, each 4-5 seconds long. The noise corpus was collected in a large variety of indoor as well as outdoor conditions.

We augment training data with above real-world noise samples. We also conduct tests on a very large repository of unseen noise resources from public noise data collection including other internal resources, MUSAN [27], and, noise data harvested from Youtube resources. For this work we didn't include RIRs in any of our simulation work. Our data simulation also includes audio gain perturbation to make the system robust to attenuated speech. We train ASR models for near loss-less decoding with frame-skips. We use 160-dim features for ASR training, where we stack the current 80-

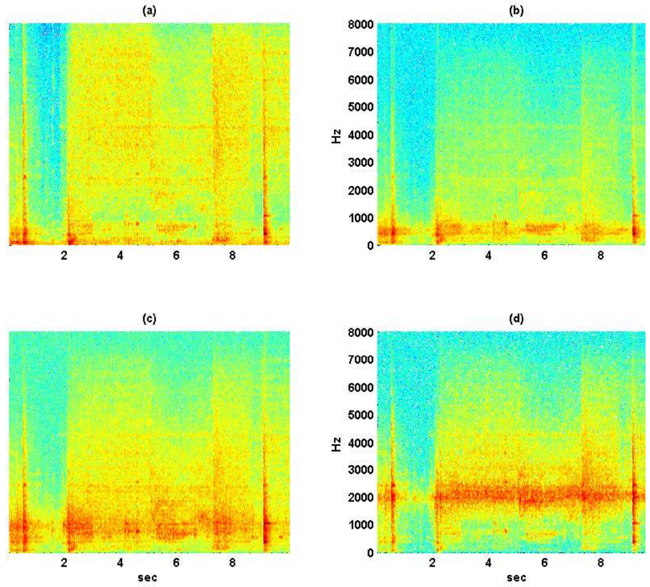

Fig. 2. *Representative subband noise samples. (a) snippet of a real-world noise. And subsequently bandpass noise samples for $B = 300$ and center frequency C as (b) 500-Hz, (c) 1000-Hz, (d) 2000-Hz.*

Table 1. *WER [%] on large scale evaluation with over 100K utts.*

Task	Baseline	Spec Aug.	Stand. Sim.	Bandpass Sim.	Purely Subband
100k Utts	22.4	22.0	19.5	18.9	19.1

dim log-Mel features to the preceding feature we skipped in model evaluation. The feature processing time window is 25-msec with 10-msec window shift. Our LSTM cells have 1024 memory units with 9k acoustic output states. We test our work on a large variety of tasks with 5-20 hrs of speech in most scenarios. We use a 5-gram language model with a vocabulary of over 1M words.

We report word error rate (WER) results on a large variety of tasks seen as well as unseen in training. More specifically, "seen" refers to tasks like Xbox that was present in our training. We also clarify that "seen" only refers to the acoustic scenario and not the specific audio data, and reiterate that training and test sets are completely distinct. Similarly "unseen" refers to completely new acoustic applications that wasn't present in training. Our broad focus is to improve ASR for unseen scenarios as well as scenarios with limited training data.

4.1. Baseline for simulation techniques

We evaluate on a large scale task with over 100k utterances in Table 1. The task spans a variety of "seen" as well as "unseen" scenarios. The task spans a very large testing across a mix of the scenarios we discussed in sec. 2. We report the Baseline, *i.e.* w/o noise augmentation, Standard Simulation, SpecAugment, and the proposed Bandpass Simulation. We also present a controlled study in "Purely Subband" Simulation, where we applied data augmentation from only the generated subband noise, *i.e.* we didn't use the original noise corpus for augmentation.

4.1.1. Standard Simulation

As expected we see a strong evidence for data augmentation work with significant WER reduction from 22.4% (w/o augmentation) to 19.5% with Standard simulation technique. Thus Standard simulation constitutes an effective reference for our work, and our merit lies in improvements over the Standard simulation.

4.1.2. SpecAugment

SpecAugment is a simple and computationally cheap simulation, it randomly masks audio spectrum along time and frequency axes. It doesn't require additional resources like noise or RIR samples. In our implementation, we apply SpecAugment to 1-sec audio segments. We randomly selected different configuration for different segments of audio for significant diversity. SpecAugment work shows minimal gains over baseline in Table 1. And more importantly it's significantly weaker than Standard simulation. Recently SpecAugment has also been developed for larger scale E2E models in [28]. However, our current evaluations suggest limited SpecAugment gains for hybrid models.

4.1.3. Bandpass Simulation

In Table 1 we also note that Bandpass simulation improves WER to 18.9%, and equivalently 3% WERR over Standard simulation on the very large scale 100k utterances task. These gains are very significant over a prior technique in SpecAugment. We also see merit in Purely-subband simulation, that improves WER to 19.1%, and is already better than our best reference in Standard simulation. Above study provides a key message that even though Purely-subband data augmentation wasn't directly trained on the original "fullband" noise corpus, it still shows better WER than Standard simulation. This suggests that training from the constituent subband noises ingests better robustness than training on original "fullband" noise corpus. Next we report individual scenarios and demonstrate significant robustness and generalization to unseen data.

4.2. Robustness to unseen speech scenarios

ASR is always evolving with new far-field devices or new acoustic conditions. We can't always train ASR for those devices to improve those scenarios. The upcoming results demonstrate broader ASR robustness to data or scenarios not seen in training. We report progress in 2 specific directions, (a) far-field device not seen in training, (b) unseen car acoustic scenario. Table 2 documents 1.5% WERR for test data from a far-field device that's seen in training, and more importantly 7.2% WERR for the unseen far-field device. Similarly, we report 7.3% WERR for the unseen car acoustic that consists of a variety of car noises in highway noise, wind noise and also at slow and fast speeds. We also achieved WER parity for near-field devices seen in training. Overall our work is applicable to existing as well as emerging devices.

4.3. Gains for emerging unseen applications

We expand this study to demonstrate gains for new ASR applications and report significant findings. In particular we report over 11.5% WERR for kids speech in Table 3 by improving WER from

Table 2. *WER [%] on seen and unseen acoustic scenarios.*

Tasks	Baseline	Stand. Sim.	Bandpass Sim.	WERR [%] - Bandpass over Stand.
Far-field - A	19.8	18.7	18.5	1.5
Far-field - B (Unseen)	20.2	16.7	15.5	7.2
Near-field	11.2	9.9	9.8	1.0
Car Noise (Unseen)	24.7	19.2	17.8	7.3

Table 3. *WER [%] on new ASR applications.*

Tasks	Baseline	Standard Sim.	Bandpass Sim.
Kids Speech	31.3	24.4	21.6
New App. - C	14.6	14.6	14.0
EN-Dialects	26.9	24.4	24.0

24.4% for Standard simulation to 21.6%. That improves the overall WERR over baseline from 22% for Standard simulation to 31% in our work. We note that, in comparison all the tasks we evaluated, Standard simulation shows significantly larger gains for kids speech, and the trend holds for Bandpass simulation as well. Kids speech is a challenging task due to evolving nature of their speech. Our training data has very limited kids speech as it's significantly hard to collect kids data at scale for ASR training, and perhaps above factors explain a larger role for data augmentation for kids speech.

We also report a new application "C", where we improve WER from 14.6% to 14.0%, a 4% WERR. We also present a study on English-dialects, where WER improves from 24.4% to 24%. Improving unified ASR for a variety of English dialects is an important objective. We see some improvements and expect future work for additional gains.

4.4. Robustness to synthesized noisy data

In Table 4 we report a study on digitally adding unseen noise to Near-field task at SNRs 10-15 dB. The noise data is from a large scale noise corpus including our internal resources, MUSAN [27], and resources from Youtube. We note a strong merit in Bandpass simulation, it improves WER from 17.8% for Standard simulation to 16.5%, a 7.3% WERR. We continued the evaluation on bandpass noise samples generated from test noise corpus. There we chose $B \in \{150, 500, 1000\}$ to avoid a direct match with (1) in training. As expected we report larger gains over Standard simulation, improving WER from 17.1% to 14.6%. In Sec. 2 we also noted audio capturing devices that may record audio in 8-kHz. Much of our prior evaluation was conducted on 16-kHz audio. In Table 4 we downsampled data to 8-kHz and report additional merit in our data simulation technique.

5. DISCUSSION

We also analyzed a few audio examples where our work resulted in lower WER. Fig. 3 illustrates representative examples. Fig. 3 (a) shows speech in presence of a complex and vibrant background noise. Fig. 3 (b) includes background noise from knocking on a door. It's satisfying to note that our work benefited those real-world usage examples, and shows promising generalization from our approach.

Table 4. *WER [%] on specific studies with synthesized data.*

Tasks	Baseline	Standard Sim.	Bandpass Sim.
Near-field	11.2	9.9	9.8
+ Noise	24	17.8	16.5
+ Subband Noise	21.6	17.1	14.6
+ Downsample (8-kHz)	25.2	15.0	13.9

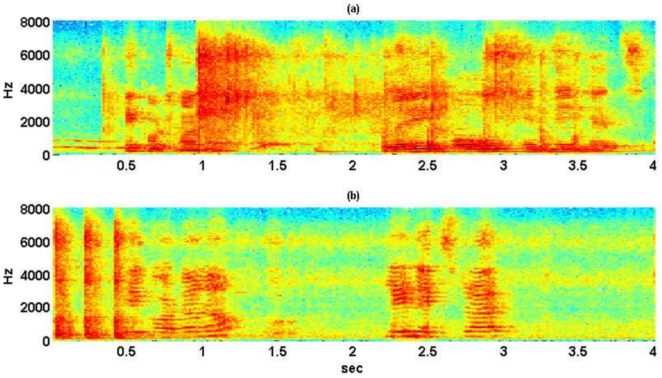

Fig. 3. *Speech spectogram for representative utterances that improved from bandpass simulation work.*

A conceptual difference between our work and SpecAugment in [22] is that our techniques closely mimics physical world and related noise sources. Our work additionally benefits from novel bandpass filters that are expected to model and reflect ambient noises present in diverse acoustic scenarios. Furthermore, SpecAugment may be unsuitable for conventional hybrid DNN-HMM approaches that's trained at frame-level. There, completely nullifying acoustic frames and mapping them to different acoustic senone targets can lead to training issues.

A future direction of this work can consider additional focus on bandpass filter designs. Besides bandpass filters, we can also design filters with a variety of frequency responses. The choice of the parameters like center frequency and filter bandwidth are also natural considerations.

6. CONCLUSION

In this work we targeted acoustic robustness to new ASR applications that may not be adequately represented in training. A large scale ASR service may not have prior information about specific ASR application, so robustness and generalization of ASR accuracy across large use cases is critical. We proposed data augmentation from subband noise samples and demonstrated significant gains over a much stronger reference in Standard data simulation. We noted 7% WERR for far-field acoustic conditions not seen in training. We also improved kids speech by 11% relative. We continued to demonstrate merit on new acoustic applications and narrowband 8-kHz data.

7. REFERENCES

[1] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[2] H. Sak, A. Senior, and F. Beaufays, "Long short-term memory recurrent neural network architectures for large scale acoustic modeling," in *Proc. Interspeech*, 2014, pp. 338–342.

[3] H. Sak, A. Senior, K. Rao, and F. Beaufays, "Fast and accurate recurrent neural network acoustic models for speech recognition," in *Proc. Interspeech*, 2015.

[4] K. Kumar, C. Liu, Y. Gong, and J. Wu, "1-D row-convolution LSTM: Fast streaming ASR at accuracy parity with LC-BLSTM," in *Interspeech*, 2020.

[5] Geoffrey Hinton, Li Deng, Dong Yu, George E Dahl, Abdelrahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al., "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal processing magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[6] Michael L Seltzer, Dong Yu, and Yongqiang Wang, "An investigation of deep neural networks for noise robust speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7398–7402.

[7] Li Deng, Jinyu Li, Jui-Ting Huang, Kaisheng Yao, Dong Yu, Frank Seide, Michael L. Seltzer, Geoffrey Zweig, Xiaodong He, Jason D. Williams, Yifan Gong, and Alex Acero, "Recent advances in deep learning for speech research at microsoft," *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 8604–8608, 2013.

[8] G.E. Dahl, D. Yu, L. Deng, and A. Acero, "Context-dependent pre-trained deep neural networks for large-vocabulary speech recognition," *IEEE Transactions on Audio, Speech, and Language Processing,*, vol. 20, no. 1, pp. 30–42, Jan. 2012.

[9] Hasim Sak, Matt Shannon, Kanishka Rao, and Françoise Beaufays, "Recurrent neural aligner: An encoder-decoder neural network model for sequence to sequence mapping," in *Proc. of Interspeech*, 2017.

[10] Dong Yu, Kaisheng Yao, Hang Su, Gang Li, and Frank Seide, "KL-divergence regularized deep neural network adaptation for improved large vocabulary speech recognition," in *Proc. ICASSP*, 2013, pp. 7893–7897.

[11] J. Xue, J. Li, D. Yu, M. Seltzer, and Y. Gong, "Singular value decomposition based low-footprint speaker adaptation and personalization for deep neural network," in *Proc. ICASSP*, 2014, pp. 6359 – 6363.

[12] K. Kumar, C. Liu, K. Yao, and Y. Gong, "Intermediate-layer dnn adaptation for offine and session-based iterative speaker adaptation," in *Interspeech*, 2015.

[13] Y. Miao, H. Zhang, and F. Metze, "Towards speaker adaptive training of deep neural network acoustic models," 2014.

[14] Xiaofei Wang, Ruizhi Li, and Hynek Hermansky, "Stream attention for distributed multi-microphone speech recognition," *Proc. Interspeech 2018*, pp. 3033–3037, 2018.

[15] Cha Zhang, Dinei Florêncio, Demba E Ba, and Zhengyou Zhang, "Maximum likelihood sound source localization and beamforming for directional microphone arrays in distributed meetings," *IEEE Transactions on Multimedia*, vol. 10, no. 3, pp. 538–548, 2008.

[16] Jonathan G Fiscus, "A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover)," in *Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on*. IEEE, 1997, pp. 347–354.

[17] K. Kumar, T. Anastasakos, and Y. Gong, "Word characters and phone pronunciation embedding for ASR confidence classifier," in *Proc. ICASSP*, 2019.

[18] Björn Hoffmeister, Ralf Schlüter, and Hermann Ney, "iCNC and iROVER: The limits of improving system combination with classification?," in *Ninth Annual Conference of the International Speech Communication Association*, 2008.

[19] Jinyu Li, Li Deng, Yifan Gong, and Reinhold Haeb-Umbach, "An overview of noise-robust automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 4, pp. 745–777, 2014.

[20] Anton Ragni, Katherine Mary Knill, Shakti P Rath, and Mark John Gales, "Data augmentation for low resource languages," 2014.

[21] Navdeep Jaitly and Geoffrey E Hinton, "Vocal tract length perturbation (vtlp) improves speech recognition," in *Proc. ICML Workshop on Deep Learning for Audio, Speech and Language*, 2013, vol. 117.

[22] Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.

[23] Chanwoo Kim, Ananya Misra, Kean Chin, Thad Hughes, Arun Narayanan, Tara Sainath, and Michiel Bacchiani, "Generation of large-scale simulated utterances in virtual rooms to train deep-neural networks for far-field speech recognition in google home," 2017.

[24] Tom Ko, Vijayaditya Peddinti, Daniel Povey, and Sanjeev Khudanpur, "Audio augmentation for speech recognition," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[25] Bo Li, Tara N Sainath, Arun Narayanan, Joe Caroselli, Michiel Bacchiani, Ananya Misra, Izhak Shafran, Hasim Sak, Golan Pundak, Kean K Chin, et al., "Acoustic modeling for google home.," 2017.

[26] Hu Hu, Tian Tan, and Yanmin Qian, "Generative adversarial networks based data augmentation for noise robust speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5044–5048.

[27] David Snyder, Guoguo Chen, and Daniel Povey, "MUSAN: A music, speech, and noise corpus," *CoRR*, vol. abs/1510.08484, 2015.

[28] D. S. Park, Y. Zhang, C. Chiu, Y. Chen, B. Li, W. Chan, Q. V. Le, and Y. Wu, "Specaugment on large scale datasets," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2020, pp. 6879–6883.