

SCENARIO AWARE SPEECH RECOGNITION: ADVANCEMENTS FOR APOLLO FEARLESS STEPS & CHIME-4 CORPORA

Szu-Jui Chen, Wei Xia, John H.L. Hansen

Center for Robust Speech Systems (CRSS), University of Texas at Dallas, TX 75080
{szujui.chen,wei.xia,john.hansen}@utdallas.edu

ABSTRACT

In this study, we propose to investigate triplet loss for the purpose of an alternative feature representation for ASR. We consider a general non-semantic speech representation, which is trained with a self-supervised criteria based on triplet loss called TRILL, for acoustic modeling to represent the acoustic characteristics of each audio. This strategy is then applied to the CHiME-4 corpus and CRSS-UTDallas Fearless Steps Corpus, with emphasis on the 100-hour challenge corpus which consists of 5 selected NASA Apollo-11 channels. An analysis of the extracted embeddings provides the foundation needed to characterize training utterances into distinct groups based on acoustic distinguishing properties. Moreover, we also demonstrate that triplet-loss based embedding performs better than i-Vector in acoustic modeling, confirming that the triplet loss is more effective than a speaker feature. With additional techniques such as pronunciation and silence probability modeling, plus multi-style training, we achieve a +5.42% and +3.18% relative WER improvement for the development and evaluation sets of the Fearless Steps Corpus. To explore generalization, we further test the same technique on the 1 channel track of CHiME-4 and observe a +11.90% relative WER improvement for real test data.

Index Terms— speech recognition, scenario aware, speech representation

1. INTRODUCTION

Significant progresses in automatic speech recognition (ASR) have taken place in recent years. Today, ASR systems are utilized in our daily lives, where a diverse range of recognition scenarios that contain distinct background acoustic conditions are observed. However, modern ASR systems are still struggling to effectively overcome noise levels and adverse background conditions, leading to unsatisfactory recognition results in daily use. It is suggested that this could be caused by less effective ASR acoustic modeling based on Mel-Frequency Cepstral Coefficients (MFCC) or log-mel filterbanks (FBANK) energies. Such features are sensitive to

noise [1], and these systems are generally trained for a specific use case and sensitive to test mismatch. Performance can further degrade for distant talking situations, where signal energy is lower, reverberation is possible, and environment signal-to-noise ratio (SNR) is lower. As a result, it is necessary to create a solution to minimize the influences from changing background acoustic conditions.

In the past, methods have been proposed to address the problem of noisy speech recognition [2]. Most focus on feature enhancement [3], or model adaptation [4]. One proposed method is based on factor-aware training. Such a technique introduces factors including noise [5], speaker [6], and/or room characteristics [7] into the training of deep neural networks (DNN) as auxiliary information. This added supporting information serves as a factor-dependent bias to the DNN which causes the output of the DNN to depend on the individual factor values. The most well-known example is the i-Vector that was originally proposed for speaker recognition [8]. Here, it is possible for us to apply it as speaker and channel representations in factor aware training.

To address diversity in acoustic characteristics, we propose adding a feature to model the acoustic characteristics, such as channel distortions and environmental noise types, in the audio. The goal here is to make the acoustic model aware of this available information, which can be summarized as a "scenario" that exists in the audio. This idea needs either several good representations for each classifiable factor, or an exceptional representation that can suitably distinguish a specific acoustic context.

Past studies have explored triplet loss as a means for improving speech technology, specially for speaker ID [9, 10]. However, to the best of our knowledge, triplet loss studies have not been explored in acoustic modeling for ASR. In this study, as motivated by past efforts in speaker recognition [9, 10], we employ a triplet-loss based representation generated by TRIPlet Loss network (TRILL) [11] for speech recognition. In that network, a subset of the AudioSet [12] that possesses the speech label is used for training in a self-supervised manner. Since the AudioSet corpus is a large dataset for general audio machine learning with general audio speech tags, it is useful due to size and scope. As a result, the triplet-loss based representation is expected to learn generalization for

This project was funded, in part, by NSF-CISE Award 2016725, and partially by the University of Texas at Dallas from the Distinguished University Chair in Telecommunications Engineering held by J. Hansen.

audio. The technique developed in [13] was used to allow the network the ability to represent segments in audio that are closer in time to be closer in the embedding space. Details are presented in Sec.3.2.

The proposed method is assessed using two datasets, the 100-hour challenge corpus of the CRSS-UTDallas Fearless Steps Corpus (Sec.4.1) and CHiME-4 corpus (Sec.4.2). Systems development employs the Kaldi speech recognition toolkit [14], and uses the same feature extraction pipeline shown in Fig.1. For the CRSS-UTDallas Fearless Steps task, we utilize a factorized time delay neural network (TDNN-f) [15] for acoustic modeling while for CHiME-4, we focus on the 1 channel track task as they employed in [16].

2. RELATED WORK

Historically, many approaches have been proposed to address noise robustness in ASR systems [17, 18]. In [5], an approach based on noise-aware training which incorporates information about the environment was introduced into DNN training. In [19], three extraction models for speaker, phone, and environment were considered, along with a multi-task joint training architecture. In [20], the invariant representation learning technique was proposed, which demonstrated significant reduction in character error rate and robustness for out-of-domain noise settings. In [21], a simple method was considered to extract a noise vector for acoustic model training. It is suggested that the technique could also be applied in online ASR by estimating the mean vector with frame-level maximum likelihood.

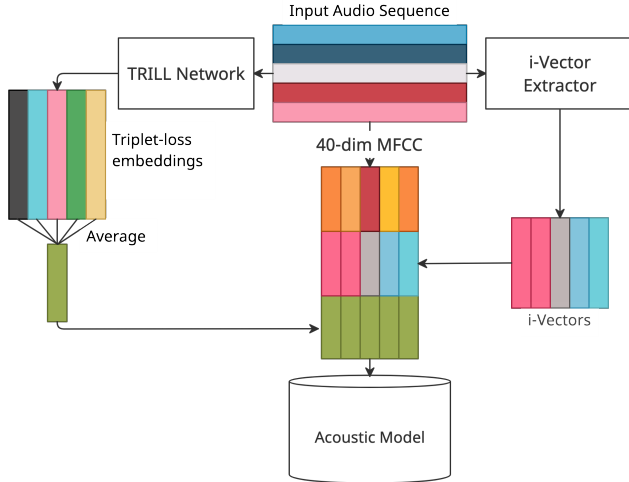


Fig. 1: Feature generation flow chart

3. PROPOSED SYSTEM

Given the challenges in robustness for ASR with CRSS-UTDallas Fearless Steps and CHiME-4, this section presents the formulation of our scenario aware based acoustic modeling to address environmental variability.

3.1. Scenario Aware

Factor aware training has been shown to be effective in ASR system development [5, 19, 21]. This training strategy produces a system that is more robust to factors such as noise, speaker, and room characteristics. Most earlier studies have used a representation for each specific distortion factor, where the extracted representations are either fed into the input layer, the hidden layer, or the output layer. In our study, we use a single representation to characterize all factors/acoustic info within an audio, including speech, which leads to a scenario aware training for the resulting acoustic model.

The input feature for our acoustic model contains two types of vectors, the first is the commonly used MFCC, along with the i-Vector which we denote as $m \in R^{1 \times M}$. The second feature is the triplet-loss vector from TRILL, which we denote as $n \in R^{1 \times N}$. This total input vector V is represented as the concatenation of M dimensional vector m and N dimensional vector n :

$$V = [m; n] \in R^{1 \times (M+N)} \quad (1)$$

Note that we average the triplet-loss embeddings over time for an audio input to form a triplet-loss vector, that is one vector for the entire audio. A flow diagram is shown in Fig 1.

3.2. Triplet-loss based Representation

Previously, the triplet-loss based representation generated by TRILL model was introduced in [11] and originally used for non-semantic downstream tasks. The pre-trained model¹ we used was trained on a subset of AudioSet [12] training set clips that own the speech label and with the ResNet-50 architecture discussed in [22], followed by a 512-dimensional embedding layer. Next, the temporal proximity is used as a self-supervision signal [13]. The idea behind using the temporal proximity is that sounds in a given environment are usually restricted to a subset of sound creating objects that are often closely related. Hence, a pair of events in the same audio should have a higher probability of being the same, or at least related in a certain level than any two audio clips randomly chosen from a large audio collection.

The collection of audio for training the triplet-loss model such as TRILL, can be depicted as a sequence of spectrogram context windows $X = [x_1, x_2, \dots, x_N]$, where $x_i \in R^{F \times T}$ with F and T represent frequency and time window. The goal for this model is to learn a map $g : R^{F \times T} \rightarrow R^D$ that transform x_i into d -dimensional space such that $\|g(x_i) - g(x_j)\| \leq \|g(x_i) - g(x_k)\|$ when $|i - j| \leq |i - k|$. This is achieved by first sampling in X a great amount of triplets $s = [x_i, x_j, x_k]$, which are known as the anchor, positive, and negative respectively, where $|i - j| \leq \tau$ and $|i - k| > \tau$ with a reasonable

¹<https://tfhub.dev/google/nonsemantic-speech-benchmark/trill/3>

time scale τ . Next, we train the model with the triplet loss:

$$\mathcal{L}(s) = \sum_{i=1}^N [\|g(x_i) - g(x_j)\|_2^2 - \|g(x_i) - g(x_k)\|_2^2 + \delta]_+ \quad (2)$$

where $\|\cdot\|$ is the L_2 norm, δ is a non-negative margin hyper-parameter, and $[\cdot]_+$ is the hinge loss. It is clear the loss will be exactly zero if all the training triplets satisfy the inequality:

$$\|g(x_i) - g(x_j)\|_2^2 + \delta \leq \|g(x_i) - g(x_k)\|_2^2 \quad (3)$$

In [11], the training task is based on a pair-wise data set with the same clip (x_i, x_j) /different clip (x_i, x_k) discrimination achieved by setting the utterance value τ to 10 seconds, which is the maximum duration of the clips in AudioSet. This makes the triplet-loss model capable of mapping audio clips with close acoustic information into similar embeddings. With the enormous size and scope of labels in AudioSet, we employed triplet-loss based representation to model the environment scenario in the audio.

4. EXPERIMENTS

The method for extracting the input features for the following experiments are all the same. Firstly, we extract 512-dimensional embeddings through the TRILL model and average these embeddings over time to form a triplet-loss vector. Next, the vector is combined with 40 dimensional MFCC features and 100 dimensional i-Vector as the input vector for acoustic model training. The feature generation pipeline is also shown in the Fig 1.

4.1. Fearless Steps Corpus Experiments

4.1.1. Data

The Fearless Steps Corpus [23] consists of 19,000 hours of actual Apollo conversational speech across 30 time-synchronized channels, with Channel#1 representing the time-synchronized IRIG timecode. The audio represents actual communications from the Apollo-11 mission including all Mission Specialists, Astronauts, and support staff over the 7-day mission to the moon. These communication channel loops have distinct acoustic characteristics (e.g., noise, distortion, background interference, etc.) from NASA analog cables to the SoundScriber recording platform with channel and system noise that contributes to loss in ASR system performance. The variability across channel loops is due to the extensive cabling, headsets, relays, etc. necessary to bridge 600 NASA specialists in different locations to allow them to communicate and work collaboratively to achieve a successful mission. All audio was recorded on 30-track analog 1 inch reel-to-reel recording tapes, and then digitized by CRSS-UTDallas initially at a 44.1kHz and later down-sampled to 8kHz, with 30 min. per data chunk for speech analysis.

For this study, we employ only the 100-hour Fearless Steps challenge corpus [24] that consists of 5 selected channels with labeled data. This includes Network Controller (NTWK), Electrical, Environmental and Consumables Manager (EECOM), Guidance Navigation and Control (GNC), Flight Director (FD), Mission Operations Control Room (MOCR). Here, we use the ASR track2 in the challenge corpus where the audio is already segmented with utterance level transcriptions. The training set is roughly 28 hours, with development set being 7.6 hours, and the evaluation set being 10.6 hours. The training set is used for both the acoustic and language model. We use the development set for computing the perplexity in language model training, and the evaluation set is used only for test.

4.1.2. Baseline System

For lexicon model, we employ the CMU dictionary² as a basic pronunciation dictionary. However, since many words in the Fearless Steps Corpus are NASA space related and not present in the CMU dictionary, we use the Phonetisaurus G2P [25] to generate pronunciations for these out-of-vocabulary words. A speaker adapted HMM-GMM is first trained on the training set to generate phoneme to audio alignments for DNN training. The TDNN-f [15] with 15 1024-dimensional layers factorized with 160-dimensional linear bottlenecks is used for acoustic modeling on the same dataset. For the language model, a basic 3-gram model was used, with pronunciation and silence probability modeling as described in [26].

4.1.3. Results and Analysis

The purpose of using a triplet-loss based representation, is to model the acoustic condition of each audio context. As shown in Fig.2, the extracted triplet-loss vectors from training data are categorized into a few different blocks. Each block can be treated as a distinct acoustic characteristic. Since we assume that the channel number corresponds to each utterance is not known, the speaker information is used instead as the label. Most speakers have utterances spread out across multiple blocks, but they do not necessarily cover every block. The triplet-loss based representation allowed for the analysis of environment scenario as shown in Fig.3, which shows selected channels that have different characteristics. This figure used a randomly selected 360-hour subset of the complete 9,000-hour Fearless Steps Corpus that contains channel number information for each audio stream. Each point represents a triplet-loss vector extracted from TRILL model from a 15 second block of audio cut from the original 30 min. sequential audio chunks.

In Table 1, word error rates (WER) are shown for experiments based on the Fearless Steps Corpus. The first row is the baseline system described in Sec.4.1.2. We found that by increasing the layer dimension of the TDNN-f to 1536 (No.2),

²<http://www.speech.cs.cmu.edu/cgi-bin/cmudict>

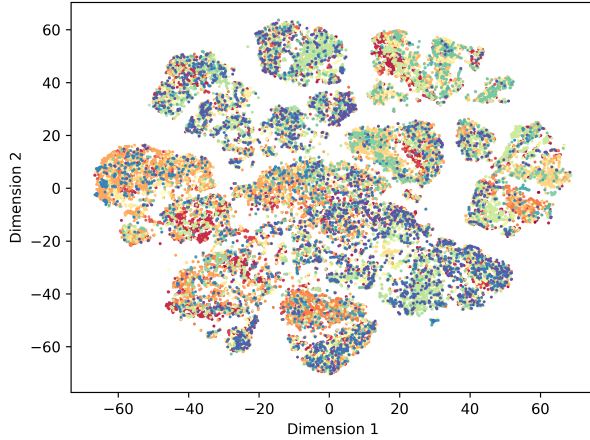


Fig. 2: t-SNE of triplet-loss vectors based on TRILL model from Fearless Steps 100-hour challenge Corpus training set. The plot includes 35k utterances (points) with each color corresponding to one of 256 Apollo-11 speakers.

Table 1: WER of Fearless Steps Corpus experiments. T-REP stands for triplet-loss based representation. *Dim* is the layer dimension of TDNN-f.

(No.) Model	i-vector	<i>Dim</i>	Dev (%)	Eval (%)
(1) Baseline	yes	1024	27.66	29.89
(2) Baseline_1536	yes	1536	27.28	29.48
(3) + T-REP matrix*	yes	1536	27.07	29.64
(4) + T-REP	no	1536	26.68	29.30
(5) + T-REP	yes	1536	26.49	29.17
(6) + T-REP & multi-style	yes	1536	26.16	28.94
Gorin et al.[27]	yes	1024	28.60	31.4

*This means we do not average over time on the embeddings.

can further reduce WER, but increasing to larger dimensions such as 2136 (not shown here), caused a loss in performance. After adding the triplet-loss based representation, we observe a 2.9% and 1.1% relative WER improvement in development and evaluation set (No.5 vs. No.2). Another discovery is that using an average over time with triplet-loss embeddings advances the WER (No.5 vs. No.3). No.4 is the only experiment without the i-Vector. This shows that MFCC with the triplet-loss based representation (No.4) is better than MFCC with i-Vector (No.2). In all, our best system (No.6) with the triplet-loss based representation and multi-style training achieves 5.42% and 3.18% relative improvement on WER in development and evaluation set respectively. The multi-style training is accomplished by adding data augmented with the room impulse response (RIR) and MUSAN corpus (music, speech, and noise). With the original data included, we are able to expand the training set size by 5x vs. the original data. Note that simply adding triplet-loss based representation provides more improvement than multi-style training (No.5 - No.2 vs. No.6 - No.5). We include the last row as a comparison to the best system in Fearless Steps Challenge Phase

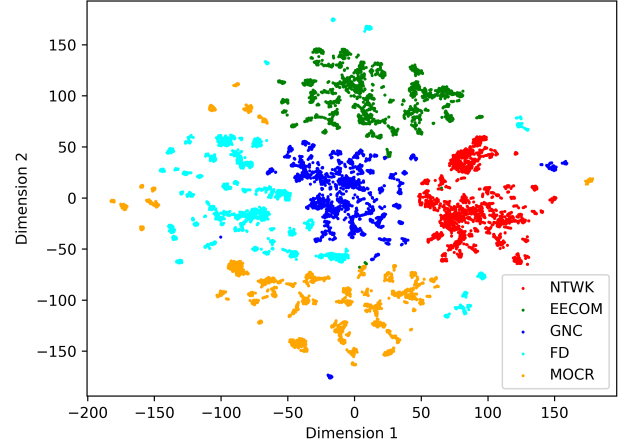


Fig. 3: t-SNE of triplet-loss vectors based on TRILL model from a randomly selected 360-hour subset of Fearless Steps Apollo-11 corpus, with 5 of 30 possible channels analyzed with data being: NTWK:66hrs, EECOM:66hrs, GNC:66hrs, FD:81hrs, MOCR:81hrs.

II, with the matched condition in both acoustic and language model and training data used in our system.

4.2. CHiME-4 corpus Experiments

4.2.1. Data

The CHiME-4 data [28] includes real data recorded in real-world noisy environments, and simulated data, that is artificially created using clean speech data mixed with noisy background data. Five locations (i.e. booth (BTH), on the bus (BUS), cafe (CAF), pedestrian area (PED), and street junction (STR)) are chosen for real data recording. The BTH recordings are used for generating the simulated data, while all the rest are for ASR evaluation.

4.2.2. Baseline System

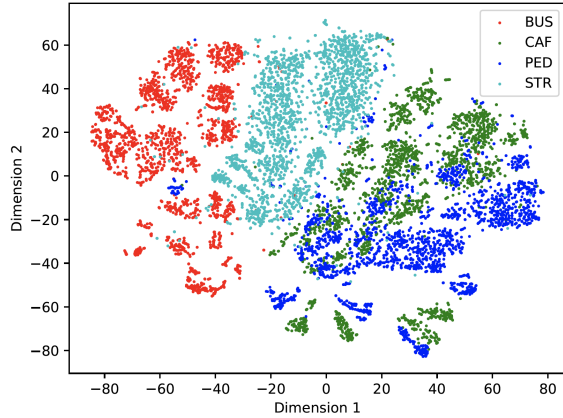
For a fair comparison, we only focus on the single channel track in the CHiME-4 challenge. The baseline system follows the work in [16], which uses a TDNN LF-MMI training on all 6 channels data and a LSTMML trained with Kaldi-RNNLM [29] on a 3-fold texts of training data. The pronunciation dictionary was also based on the CMU dictionary.

Table 2: WER of CHiME-4 1 channel track experiments. T-REP stands for triplet-loss based representation.

Model	Dev (%)		Test (%)	
	real	simu	real	simu
Baseline	5.28	6.52	11.68	11.98
+ T-REP	4.43	6.55	10.29	12.03

Table 3: WER by location of the test set in CHiME-4 1 channel track experiments. T-REP stands for triplet-loss representation.

Model	real (%)				simu (%)			
	BUS	CAFE	PED	STR	BUS	CAFE	PED	STR
Baseline	18.81	11.88	8.46	7.56	8.72	13.80	11.49	13.90
+ T-REP	15.40	11.3	8.09	6.37	8.44	14.66	11.56	13.45
relative improvement	18.13	4.88	4.37	15.74	3.21	-6.23	-0.61	3.24

**Fig. 4:** t-SNE of triplet-loss vectors based on TRILL model from CHiME-4 Corpus training set. The plot includes 8738 utterances (points).

4.2.3. Results and Analysis

In Table 2, we demonstrate the effect of adding the triplet-loss based representation. We observe a 16.10% and 11.90% relative WER improvement in real development and real test data by adding the triplet-loss based representation with only a small loss in simulation data.

The Fig.4 shows the t-SNE plot of triplet-loss vectors extracted from training set of CHiME-4 corpus. We can see a clear separation of each location, where only the CAF and PED are more overlapped than others. This observation matches the WER improvement. Also, it is suggested that this is one of the reasons why CHiME-4 benefits more from triplet-loss based representation versus the Fearless Steps Corpus.

We further investigate the effectiveness of triplet-loss based representation using the environment difference in the test set. In Table 3, we show a greater improvement for real data versus simulation data. With the observation in Fig.4 in mind, it is shown that BUS and STR environment locations have greater improvement than CAFE and PED environment locations. This leads to the conclusion that the more distinct the acoustic context is in the audio from others, the more beneficial triplet-loss based representation will help.

4.3. Analysis on Triplet-loss Representation Performance

We note that there is a wide gap in performance of triplet-loss based representation between Fearless Steps and CHiME-4

corpora (i.e. 1.1% compared to 11.90% relative improvements on WER). It is suggested that this is based on the dissimilar in formation of the acoustics for the difference in system improvements. As mentioned in Sec.4.1.1, the audio for Fearless Steps are all analog recordings, where audio cable routing and channel recording conditions cause additional background noise that become a distinct characteristic of each channel. However, the specific distortion of these channels, such as strong low frequency harmonics, are highly unsophisticated, compared to the diverse general background noises in CHiME-4 data. Another possible reason is that the audio conversational-turn duration in Fearless Steps Corpus are commonly short, with a mean of 1.93 sec, and standard deviation of 3.28; while audio duration in CHiME-4 has a mean of 7.44 sec and standard deviation of 2.86. Here, 22% of the training set consists of audio turns which are less than 1 sec. These short duration make it hard for triplet-loss based representation to be as meaningful or effective.

5. CONCLUSION

This study has considered a triplet-loss approach as our proposed method for scenario aware speech recognition. To employ triplet-loss based representation, we utilize the TRILL model (Sec.3.2) to model all factors/acoustic info within an utterance, leading to a scenario aware ASR system. This technique is especially beneficial for real data when compared to simulation data. Furthermore, the more distinct the background acoustic structure is from each other, the greater the improvement possible. The system achieved 5.42% and 3.18% relative WER improvement on the development and evaluation test sets of the Fearless Steps Corpus, and 11.90% relative WER improvement on real test data of CHiME-4 corpus.

Our future work will explore alternative representations trained for different architectures and data. Also, we will further explore the integration between neural embeddings and the resulting acoustic model.

6. ACKNOWLEDGEMENTS

The authors would like to express our sincere thanks for valuable discussions with Wei-Cheng Lin, Midia Yousefi, and Aditya Joglekar.

7. REFERENCES

- [1] Utpal Bhattacharjee, Swapnanil Gogoi, and Rubi Sharma, "A statistical analysis on the impact of noise on mfcc features for speech recognition," in *2016 International Conference on Recent Advances and Innovations in Engineering (ICRAIE)*. IEEE, 2016, pp. 1–5.
- [2] Yifan Gong, "Speech recognition in noisy environments: A survey," *Speech communication*, vol. 16, no. 3, pp. 261–291, 1995.
- [3] Dong Yu, Li Deng, Jasha Droppo, Jian Wu, Yifan Gong, and Alex Acero, "A minimum-mean-square-error noise reduction algorithm on mel-frequency cepstra for robust speech recognition," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 4041–4044.
- [4] Michael L Seltzer, Alex Acero, and Kaustubh Kalgankar, "Acoustic model adaptation via linear spline interpolation for robust speech recognition," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2010, pp. 4550–4553.
- [5] Michael L Seltzer, Dong Yu, and Yongqiang Wang, "An investigation of deep neural networks for noise robust speech recognition," in *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE, 2013, pp. 7398–7402.
- [6] Tian Tan, Yanmin Qian, Dong Yu, Souvik Kundu, Liang Lu, Khe Chai Sim, Xiong Xiao, and Yu Zhang, "Speaker-aware training of lstm-rnns for acoustic modelling," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5280–5284.
- [7] Ritwik Giri, Michael L Seltzer, Jasha Droppo, and Dong Yu, "Improving speech recognition in reverberation using a room-aware deep neural network and multi-task learning," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5014–5018.
- [8] Najim Dehak, Patrick J Kenny, Réda Dehak, Pierre Dumouchel, and Pierre Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, 2010.
- [9] Chunlei Zhang, Kazuhito Koishida, and John HL Hansen, "Text-independent speaker verification based on triplet convolutional neural network embeddings," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 9, pp. 1633–1644, 2018.
- [10] Chunlei Zhang, Fahimeh Bahmaninezhad, Shivesh Rangan, Harishchandra Dubey, Wei Xia, and John HL Hansen, "UTD-CRSS systems for 2018 NIST speaker recognition evaluation," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 5776–5780.
- [11] Joel Shor, Aren Jansen, Ronnie Maor, Oran Lang, Omry Tuval, Felix de Chaumont Quiry, Marco Tagliasacchi, Ira Shavitt, Dotan Emanuel, and Yinnon Haviv, "Towards learning a universal non-semantic representation of speech," *Proc. Interspeech 2020*, pp. 140–144, 2020.
- [12] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 776–780.
- [13] Aren Jansen, Manoj Plakal, Ratheet Pandya, Daniel PW Ellis, Shawn Hershey, Jiayang Liu, R Channing Moore, and Rif A Saurous, "Unsupervised learning of semantic audio representations," in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 126–130.
- [14] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, et al., "The kaldi speech recognition toolkit," in *IEEE 2011 workshop on automatic speech recognition and understanding*. IEEE Signal Processing Society, 2011, number CONF.
- [15] Daniel Povey, Gaofeng Cheng, Yiming Wang, Ke Li, Hainan Xu, Mahsa Yarmohammadi, and Sanjeev Khudanpur, "Semi-orthogonal low-rank matrix factorization for deep neural networks," in *Interspeech*, 2018, pp. 3743–3747.
- [16] Szu-Jui Chen, Aswin Shanmugam Subramanian, Hainan Xu, and Shinji Watanabe, "Building state-of-the-art distant speech recognition using the CHiME-4 challenge with a setup of speech enhancement baseline," *arXiv preprint arXiv:1803.10109*, 2018.
- [17] 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.
- [18] John HL Hansen, Abhishek Kumar, and Pongtep Angkitrakul, "Environment mismatch compensation using average eigenspace-based methods for robust speech

- recognition,” *International Journal of Speech Technology*, vol. 17, no. 4, pp. 353–364, 2014.
- [19] Yanmin Qian, Tian Tan, and Dong Yu, “Neural network based multi-factor aware joint training for robust speech recognition,” *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 24, no. 12, pp. 2231–2240, 2016.
- [20] Davis Liang, Zhiheng Huang, and Zachary C Lipton, “Learning noise-invariant representations for robust speech recognition,” in *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 56–63.
- [21] Desh Raj, Jesus Villalba, Daniel Povey, and Sanjeev Khudanpur, “Frustratingly easy noise-aware training of acoustic models,” *arXiv preprint arXiv:2011.02090*, 2020.
- [22] Shawn Hershey, Sourish Chaudhuri, Daniel PW Ellis, Jort F Gemmeke, Aren Jansen, R Channing Moore, Manoj Plakal, Devin Platt, Rif A Saurous, Bryan Seybold, et al., “CNN architectures for large-scale audio classification,” in *2017 IEEE international conference on acoustics, speech and signal processing (icassp)*. IEEE, 2017, pp. 131–135.
- [23] John HL Hansen, Abhijeet Sangwan, Aditya Joglekar, Ahmet Emin Bulut, Lakshmish Kaushik, and Chengzhu Yu, “Fearless steps: Apollo-11 corpus advancements for speech technologies from earth to the moon,” in *INTERSPEECH*, 2018, pp. 2758–2762.
- [24] Aditya Joglekar, John HL Hansen, Meena Chandra Shekar, and Abhijeet Sangwan, “Fearless steps challenge (FS-2): Supervised learning with massive naturalistic apollo data,” *Proc. Interspeech 2020*, pp. 2617–2621, 2020.
- [25] Josef R Novak, Nobuaki Minematsu, and Keikichi Hirose, “WFST-based grapheme-to-phoneme conversion: Open source tools for alignment, model-building and decoding,” in *Proceedings of the 10th International Workshop on Finite State Methods and Natural Language Processing*, 2012, pp. 45–49.
- [26] Guoguo Chen, Hainan Xu, Minhua Wu, Daniel Povey, and Sanjeev Khudanpur, “Pronunciation and silence probability modeling for asr,” in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.
- [27] Arseniy Gorin, Daniil Kulko, Steven Grima, and Alex Glasman, ““This is Houston. Say again, please”. the behavox system for the Apollo-11 Fearless Steps Challenge (Phase II).,” *Power*, vol. 400, no. 800, pp. 1200.
- [28] Emmanuel Vincent, Shinji Watanabe, Aditya Arie Nugraha, Jon Barker, and Ricard Marxer, “An analysis of environment, microphone and data simulation mismatches in robust speech recognition,” *Computer Speech & Language*, vol. 46, pp. 535–557, 2017.
- [29] Hainan Xu, Ke Li, Yiming Wang, Jian Wang, Shiyin Kang, Xie Chen, Daniel Povey, and Sanjeev Khudanpur, “Neural network language modeling with letter-based features and importance sampling,” in *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2018, pp. 6109–6113.