ISSN 2303-4521
Original Research

Periodicals of Engineering and Natural Sciences
Vol. 9, No. 4, November 2021, pp.921-927

Enhancement automatic speech recognition by deep neural networks

Muhammad D. Hassan [1], Ali Nejdet Nasret [2], Mohammed Rashad Baker[3], Zuhair Shakor Mahmood[4]
[1] College of Health and Medical Technology, Northern Technical University
[2] Electronic Department Kirkuk Technical Institute, Northern Technical University
[3] Department of Computer Engineering Techniques, Imam Ja'afar Al-Sadiq University
[4] Electronic Department Kirkuk Technical Institute, Northern Technical University

ABSTRACT

The performance of speech recognition tasks utilizing systems based on deep learning has improved dramatically in recent years by utilizing different deep designs and learning methodologies. A popular way to boosting the number of training data is called Data Augmentation (DA), and research shows that using DA is effective in teaching neural network models how to make invariant predictions. furthermore, EM approaches have piqued machine-learning researchers' attention as a means of improving classifier performance. In this study, have been presented a unique deep neural network speech recognition that employs both EM and DA approaches to improve the system's prediction accuracy. firstly, reveal an approach based on vocal tract length disturbance that already exists and then propose a Feature perturbation is an alternative Data Augmentation approach. in order to make amendment training data sets. This is followed by an integration of the posterior probabilities obtained from several DNN acoustic models trained on diverse datasets. The study's findings reveal that the proposed system's recognition skills have improved.

Keywords: Linear Logistic Regression, Ensemble Method, Deep Learning, Data Augmentation, Speech Recognition

Corresponding Author:

Muhammad D. Hassan
College of Health and Medical Technology
Northern Technical University
College of Health and Medical Technology, Kirkuk, Iraq
E-mail: alinajdet@ntu.edu.iq

1. Introduction

Acoustic, linguistic, and pronunciation dictionary models all go into making a voice recognition system work. Another important factor in modeling success is how much data can use to train your model. Speech data and acoustic models have been the focus of intense study in the last several years, with promising results. Voice recognition systems using deep neural networks perform better than GMM-based ones in continuous word recognition tasks and phone throughout the previous decade [1]. Recent developments in deep learning approaches like convolutional neural networks (CNN) and (RNN) deep recurrent neural networks have had a significant impact on speech recognition. Voice recognition deep learning has been praised as better than standard GMM-based systems, however in fact there aren't many big differences in performance amongst the several deep algorithms that have been created in the field [2-4]. On the other hand, it is possible that integrating these several systems into a single one is the best option [5, 6]. In addition, another area of speech recognition research was devoted to the development of effective DA approaches [7, 8]. When it comes to voice recognition, enormous amounts of transcribed training data are required, but this is not true for all languages [9-11]. DA, on the other hand, proposes using several forms of alterations to supplement the voice data. A popular method is to expand the amount of training data that is available by whatever means necessary. As an example, utilizing DNN-based acoustic modeling, on the TIMIT phoneme identification test, using VTLP helped people perform better [12, 13]. It is suggested in this research to use current methodologies to enhance the recognition performance in a new DNN-based voice recognition system. Approaches such as design analysis (DA) and experimental modeling (EM) are included into one system. DNN acoustic modeling makes good use of DA methods. For the first step, investigated the VTLP and suggest an alternative data-augmentation method based on feature perturbation. Several deep neural network acoustic technologies trained on diverse data sets use EM techniques to merge their posterior probabilities for improved prediction accuracy [14-17]. A major voting scheme, an LLR and average voting scheme for Fusion and Calibration are all tested in this study. The results of the experiments show that the suggested system works.

2. Expansion of the data

It is common to use augmentation techniques to raise the quantity of training data available. It's an essential part of image and speech recognition systems, respectively [18, 19]. A vast voice database is required to train NN in speech recognition systems [20]. DA comes in helpful when working with small data information. It is feasible to develop speech databases, which may subsequently be used to enhance the accuracy. A change in the length of the vocal tract [21]. The (VTLP) Vocal tract length perturbation phoneme recognition challenge has been enhanced by applying DNN-based audio modeling for (VTLP) vocal tract length perturbation. Because of the inclusion of LVCSR, the possibilities of VTLP were greatly expanded [22, 23]. Vocal tract length perturbation warping parameters from a small set of perturbation index were shown to be more effective [20]. A random warping factor is determined from the range 0.8,1.0 for each syllable in the training set to skew the frequency axis. This causes a little distortion in the speaker's vocal tract length, which in turn distorts the original utterance's speech spectrum and creates a new copy. The original features are duplicated three times using a set of warping factors of 0.8, 1.0, and 1.2. The warping variables are also applied uniformly throughout the training set [24].

2.1. Perturbation of a specific feature

It is our goal to compare our new approach to DA to the currently used augmentation method, known as VTLP. The goal of feature perturbation is to introduce random values into the extracted acoustic feature vectors. To create a reworked version of the source data , it is beneficial to decrease the speech quality, modify the speaker's voice, and so on. It is not uncommon for audio perturbation techniques to produce results like this. $s(n)$ may be represented as the speech signal plus the original message signal, which is written as $s(n) = m(n) + e(n)$, where $m(n)$ and $e(n)$ are the message and noise signals, respectively. The resultant transformation is $feat^s = feat^m + feat^e$, which may be achieved by using a feature extraction approach like MFCC or PLP. Our plan is to add random values $feat^r$ to the retrieved characteristics $feat^s$ for each utterance to see how they change (e.g., [0, 1]). If we look at it this way, the additional random values may be considered as an intentional noise feature. The random features are denoted by $feat^r$ while the perturbed features are denoted by $feat^p$. The greater the amplitude range, the more unrealistic the distortion will be. Features now have a twofold increase in size due to the warped ones being as large as the originals.

3. Techniques for model-based fusion

3.1. Methods used by an ensemble

When using more than one classifier technique, the goal is to provide a single classification result by integrating the predictions of several classifiers. As a whole, the ensemble's results are more accurate than the sum of its individual parts. Diverse classifiers may be obtained and combined in a variety of ways.

Either a major voting scheme or an average voting scheme is the most straightforward and economical way to combine log-likelihood results. Assuming that there are K acoustic recognition modelsR_k, the majority of voting will be done by utilizing the argmax function to combine them. There is no model in this class that can anticipate a value as severe as the one returned by this function. Using the average as a third, way to combine several classifiers is possible. The predictions of all *K* models are averaged for an input *x*, as follows:

$$\acute{Y} = \frac{1}{K} \sum_{k=1}^{K} R_k(x) \tag{1}$$

3.2. Linear logistic regression methods are used to combine and calibrate the data

One of the most popular methods for transforming a list of characteristics into probability ratios is linear logistic regression. The logistic function is the foundation of the classic logistic regression model, which is defined as:

$$P(x; \alpha, \beta) = \frac{1}{1+e^{-y(\alpha x+\beta)}} \qquad (2)$$

Where y is the class of the result and x is the feature index of the input.
This standard is reformed to take into consideration numerous recognizers when it comes to fusion and calibration. As a result, a new recognition model $\acute{R}$ was generated and represent by $R_k, = 1..k,$ where each and every model predicts the same log-likelihood in the same N classes f and where $\theta = \alpha_1, \alpha_2, \dots, \alpha_K, \acute{\beta}$
Where $\alpha_k \in \mathbb{R}$ is a value of weight in the model , $\acute{\beta} \epsilon \mathbb{R}^N$ is a log-likelihood $index$. The new recognizer's output is as follows for an input x:

$$\acute{R} = \acute{\beta} + \sum_{k=1}^{K} R_k(x) \qquad (3)$$

It is a fusion or combination when the weighted total of all the recognition models is added together. On the fusion output , the vector $\acute{b}$ is employed as an affine calibration transformation.

4. Planned approach

Two crucial features are proposed in this work to enhance standard ASR systems. For the recognition phase, various acoustic models are trained, with the best results being utilized. However, combining all of the previously trained models may prove to be more efficient. As a result, apply the score fusion and EM concepts to the problem of merging the predictions from different acoustic models. The ASR architecture is shown in Figure 1.

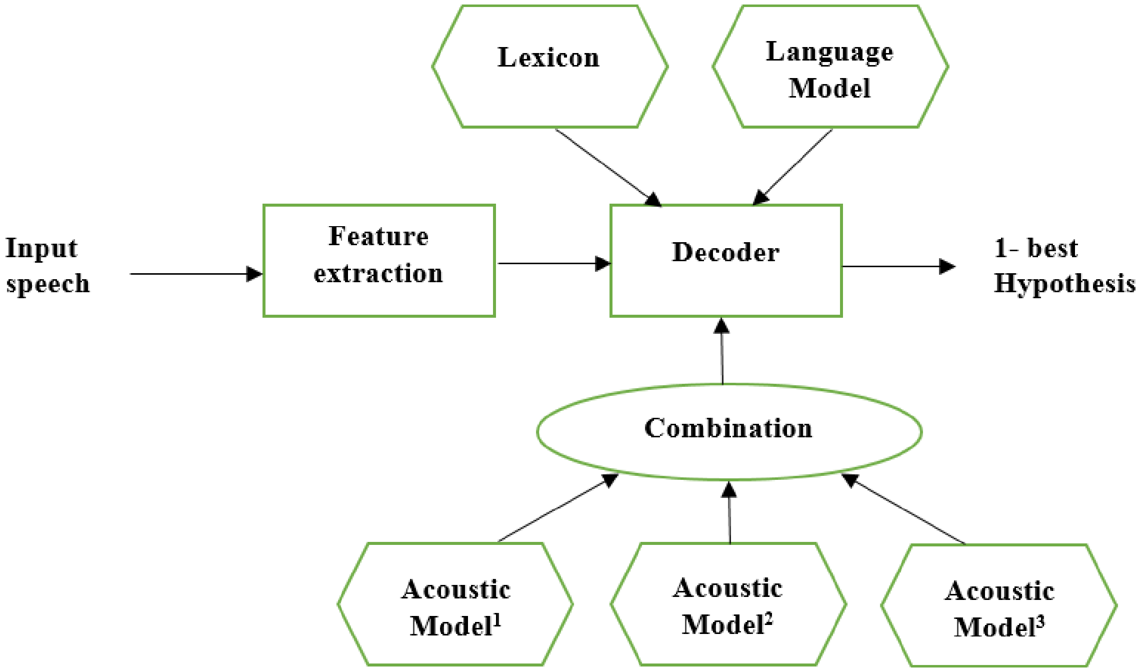

Figure 1. The ASR architecture that has been suggested

Essentially, the concept is that since each acoustic model uses a distinct set of information to describe the input signal, the models may be used in an ensemble effectively. Using distinct sets of acoustic models and different sets of data is one approach. The final log-posteriors are generated by combining the log-posteriors generated thus far. DA approaches are thus used in this study to create acoustic models from various sources of data. After that, the models are integrated with the lexicon and language and sent to the decoder to create the final form transcription.

5. Setup for experimentation

provide the results of a problem using English automated voice recognition. At this time, the database has around 34 hours of speech in it then separated the information into two groups: a training set and a validation set. Due to possible speaker overlap, the data was divided per speaker to ensure that speakers from same set would not appear in the other same sets used for training and testing purposes. Finally, 34 hours are dedicated to training, with the final exam set requiring the remaining time. In addition to the training transcripts, extra data was gathered from other sources and used to train two tri-gram language models. There was in fact a Medium Language Model (MLM) and a Large Language Model (LLM), both of which used roughly 20 million words each, as a starting point (LLM). For this project, we utilized the CMU Sphinx Speech Recognition Toolkit1 lexicon. There are 62K distinct words in the French lexicon. Aside from that, we add terms from the training corpus to our lexicon. In order to produce the pronunciations for these terms, the Sequitur G2P toolbox was used.

5.1. Features

MFCCs, which measure the mel-frequency cepstral coefficient, were employed in this study. Even yet, the gleaned features are not utilized in actual training. In fact, a number of transformations are used to enhance the extract the most important data and feature representation. Before anything else, the static feature vectors must be generated. Dimensions of a fixed vector layered with five spliced vectors 101-dimensional vector is then reduced to 39 dimensions using (LDA)Linear Discriminant Analysis, a feature dimensionality reduction technique. It is followed by applying a feature modification called the (MLLT) Maximum Likelihood Linear Transform. Interspeaker variability is then normalized using a (fMLLR) feature space Maximum Likelihood Linear Regression. The 39-dimensional characteristics needed to train the DNN models make up the final vector.

5.2. Configuration of the DNN system

Acoustic models built on DNN were used in our research. For a variety of LVCSR tasks, these models provide cutting-edge results at high speed. That is why the advantages attributable to the use of DA techniques and model fusion approaches can be verified. The following was the set-up of the neural networks in the system: The I/P features are the 39-dimensional characteristics described Previously. Using the softmax O/P layer, we may get context-dependent HMM states from the output labels by taking the log-posterior of those states (there were approx3000 states in this study). This is true for all concealed layers, regardless of how many are present.

A total of three forms of deep neural network are examined in this study: first one is DBMDNN, which employs the initialization weights from deep Boltzmann machines (DBMs) and hyperbolic tangent activation function, denoted as tanh -DNN and deep maxout networks with p-norm nonlinearity function20, termed as pnorm-DNN

6. Findings from experimentation

Different studies have been carried out utilizing various arrangements, while test and training data sets were maintained separate. When calculating the recognition process's grade, the Word Error Rate (WER) was taken into account. Performance was also assessed using the Phone Error Rate (PER). The findings are shown in Table 1 for a variety of training methods.

Table 1. Observations made during research on a dummy data collection

WER			PER	
decoding & Training	LLM	MLM	LLM	MLM
HMM-GMM	15.12	14.90	20.19	19.57
SGMM	14.47	14.32	18.35	18.17

WER			PER	
DNN-tanh 3-layers	14.77	14.40	17.54	17.15
DNN-tanh 5-layers	14.44	14.16	16.99	16.65
DNN-tanh 7-layers	14.41	14.12	16.78	16.51
DRM-DNN 3-layers	15.26	14.60	16.31	16.11
DRM-DNN 5-layers	15.04	14.77	16.15	15.992
DRM-DNN 7-layers	15.13	14.83	16.26	16.03
DNN-Pnorm 3-1ayers	14.58	14.31	16.77	16.28
DNN-Pnorm 5-1ayers	14.41	14.15	16.39.	15.88
DNN-Pnorm 7-1ayers	14.42	14.30	16.23	15.96

In this situation, DNN definitely outperforms all alternatives. Deep learning algorithms like GMM-HMM and SGMM21 outperform more conventional methods, according to these results. The performance of DNN with seven hidden layers is similarly less than that of DNN with five hidden layers. A very deep network could not be utilized to train it properly because to the little training set. If we compare DNN-tanh and DBM-DNN to deep maxout networks, the latter two perform worse. In comparison to DNN-WER pnorm's of 16.78 percent, DNN-WER tanh's of 17.50 percent and DBM-WER DNN's of 16.91 percent are equally equal. Finally, an ASR system that uses a large language model outperforms one that uses a medium one.

Table 2. Findings from the data augmentation process

PER			WER	
Training & decoding	MFCC	VTLP	MFCC	VTLP
SGMM2(original data)	14.32		19.17	
SGMM2(generated data)	14.28	13.97	18.92	18.86
DNN(original data)	15.15		16.88	
DNN(generated data)	15.04	14.36	16.85	16.84

As shown in Table 2, the results from our experiments with feature perturbation and VTLP are shown. Pnorm-DNN is being used in this study. It was found that using feature perturbation training data with the SGMM model and VTLP, the newly constructed DNN outperformed the baseline DNN by 0.2 percent in WER on the test set. DNN models, on the other hand, failed to enhance either strategy. Using a feature perturbation vs. VTLP, we found that the former was equally effective. Even with smaller feature sizes, feature perturbation outperformed VTLP in terms of performance. Last but not least, tests were carried out to see how well the proposed model fusion strategy worked in practice. Three alternative approaches of DNN fusion are used in these studies to merge the best models: average, argmax, and LLR fusion. The original features are disrupted, and the VTLP features are used. This is seen in Figure 2 during LLR fusion and calibration training.

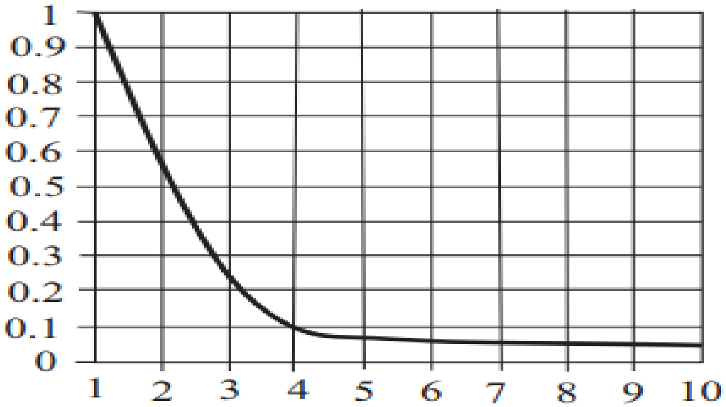

Figure 2.Training results for linear logistic regression

Table 2. Fused sound models from many sources

	PER	WER
LLR	13.09	15.40
Average	13.17	15.41
Argmax	12.59	15.37

During training, the model seems to converge quicker. As a result, we put the LLR system through its paces for a few iterations. Model fusion strategies on the test set have improved performance, as seen in Table 3. When compared to a DNN baseline, model fusion was shown to be beneficial. VTLP -DNN- features, perturbation all perform almost as well when these three models are merged. WER improves by 0.5 percent when these three DNN models are used together. A WER of 16.39% was obtained by using the LLR and average approaches, whereas the argmax strategy got an even higher WER of 16.36% on the test set. Using the PER measure, a similar rise can be seen. These findings support the hypothesis that integrating previously trained models may improve performance, and they also highlight the value of using EM approaches to boost system accuracy

7. Conclusion

A fusion architecture for voice recognition is adapted in this paper for the first time to the problem at hand. Model fusion and evolutionary modeling (EM) are both used in the suggested design. There is a novel method for enhancing voice recognition performance described in this work. Fusion Model approaches, which have been proved to be very efficient on a variety of tasks, are used in this system, as are DA techniques, which have been shown to improve ASR performance. A new audio enhancement method based on feature perturbation is presented in this study. In this method is less expensive to implement and has been shown to be as effective as the VTLP method. Finally, have been used the principles of fusion and EM to create a single ASR system that incorporates all models. The logical approach is to combine models using the average and argmax techniques. There is also an assessment of the LLR for Fusion and Calibration. The suggested system's usefulness and efficiency were shown via experiments on an ASR task in English.

References

[1] D. Baby, T. Virtanen, and J. F. Gemmeke, "Coupled dictionaries for exemplar-based speech enhancement and automatic speech recognition," *IEEE/ACM Transactions on Audio, Speech, Language Processing,* vol. 23, no. 11, pp. 1788-1799, 2015.

[2] M. Yousefi and J. H. Hansen, "Block-based high performance CNN architectures for frame-level overlapping speech detection," *IEEE/ACM Transactions on Audio, Speech, Language Processing,* vol. 29, pp. 28-40, 2020.

[3] N. Saleem, M. I. Khattak, M. Al-Hasan, and A. B. Qazi, "On learning spectral masking for single channel speech enhancement using feedforward and recurrent neural networks," *IEEE Access,* vol. 8, pp. 160581-160595, 2020.

[4] H. Salim, S. M. Najeeb, and S. M. Ali, "Finding the discriminative frequencies of motor electroencephalography signal using genetic algorithm," *TELKOMNIKA,* vol. 19, no. 1, pp. 285-291, 2021.

[5] G. Gelly and J.-L. Gauvain, "Optimization of RNN-based speech activity detection," *IEEE/ACM Transactions on Audio, Speech, Language Processing,* vol. 26, no. 3, pp. 646-656, 2017.

[6] A. Ghazi, S. Aljunid, S. Z. S. Idrus, A. Fareed, A. Al-dawoodi, Z. Hasan, R. Endut, N. Ali, A. H. Mohsin, and S. S. Abdullah, "Hybrid Dy-NFIS & RLS equalization for ZCC code in optical-CDMA over multi-mode optical fiber," *Periodicals of Engineering Natural Sciences,* vol. 9, no. 1, pp. 253-276, 2021.

[7] S. U. Wood, J. K. Stahl, and P. Mowlaee, "Binaural codebook-based speech enhancement with atomic speech presence probability," *IEEE/ACM Transactions on Audio, Speech, Language Processing,* vol. 27, no. 12, pp. 2150-2161, 2019.

[8] Z. S. Mahmood, A. N. N. Coran, and A. Y. Aewayd, "The Impact of Relay Node Deployment In Vehicle Ad Hoc Network: Reachability Enhancement Approach," in *2019 Global Conference for Advancement in Technology (GCAT)*, 2019, pp. 1-3: IEEE.

[9] A. Asaei, M. Cernak, and H. Bourlard, "Perceptual information loss due to impaired speech production," *IEEE/ACM Transactions on Audio, Speech, Language Processing,* vol. 25, no. 12, pp. 2433-2443, 2017.

[10] A. Al-zubidi, R. K. Hasoun, and H. Alrikabi, "Mobile Application to Detect Covid-19 pandemic by using Classification Techniques: Proposed System," *International Journal of Interactive Mobile Technologies,* vol. 15, no. 16, pp. 34-51, 2021.

[11] A. Ghazi, S. Aljunid, A. Fareed, S. Z. S. Idrus, C. M. Rashidi, A. Al-dawoodi, and A. M. Fakhrudeen, "Performance Analysis of ZCC-Optical-CDMA over SMF for Fiber-To-The-Home Access Network," in *Journal of Physics: Conference Series*, 2020, vol. 1529, no. 2, p. 022013: IOP Publishing.

[12] L. Chai, J. Du, Q.-F. Liu, and C.-H. Lee, "A Cross-Entropy-Guided Measure (CEGM) for Assessing Speech Recognition Performance and Optimizing DNN-Based Speech Enhancement," *IEEE/ACM Transactions on Audio, Speech, Language Processing,* vol. 29, pp. 106-117, 2020.

[13] A. Al-Dawoodi, "Neural network equalization scheme to improve channel impulse response at the receiver for optical mode division multiplexing," *Universiti Utara Malaysia,* 2016.

[14] T. A. M. Celin, G. A. Rachel, T. Nagarajan, and P. Vijayalakshmi, "A weighted speaker-specific confusion transducer-based augmentative and alternative speech communication aid for dysarthric speakers," *IEEE Transactions on Neural Systems and Rehabilitation Engineering,* vol. 27, no. 2, pp. 187-197, 2018.

[15] H. Alrikabi, and H. Tauma "Enhanced Data Security of Communication System using Combined Encryption and Steganography," *International Journal of Interactive Mobile Technologies*, vol. 15, no. 16, pp. 144-157, 2021.

[16] B. Mohammed, R. Chisab, and H. Alrikabi, "Efficient RTS and CTS Mechanism Which Save Time and System Resources," *international Journal of Interactive Mobile Technologies,* vol. 14, no. 4, pp. 204-211, 2020.

[17] A. Ghazi, S. Aljunid, S. Z. S. Idrus, R. Endut, N. Ali, A. Amphawan, A. Fareed, A. Al-dawoodi, and A. Noori, "Donut Modes in Space Wavelength Division Multiplexing: Multimode Optical Fiber Transmission based on Electrical Feedback Equalizer," in *Journal of Physics: Conference Series*, 2021, vol. 1755, no. 1, p. 012046: IOP Publishing.

[18] X. Cui, V. Goel, and B. Kingsbury, "Data augmentation for deep neural network acoustic modeling," *IEEE/ACM Transactions on Audio, Speech, Language Processing,* vol. 23, no. 9, pp. 1469-1477, 2015.

[19] S. Chen and H. Leung, "Concurrent data transmission through analog speech channel using data hiding," *IEEE Signal Processing Letters,* vol. 12, no. 8, pp. 581-584, 2005.

[20] S. Chandrakala and N. Rajeswari, "Representation learning based speech assistive system for persons with dysarthria," *IEEE Transactions on Neural Systems Rehabilitation Engineering,* vol. 25, no. 9, pp. 1510-1517, 2016.

[21] L. Baghai-Ravary and S. W. Beet, "Multistep coding of speech parameters for compression," *IEEE transactions on speech audio processing,* vol. 6, no. 5, pp. 435-444, 1998.

[22] Z. Zhou, X. Hong, G. Zhao, and M. Pietikäinen, "A compact representation of visual speech data using latent variables," *IEEE transactions on pattern analysis machine intelligence,* vol. 36, no. 1, pp. 1-1, 2013.

[23] H. Salim, and N. A. Jasim, "Design and Implementation of Smart City Applications Based on the Internet of Things," *International Journal of Interactive Mobile Technologies (iJIM),* vol. 15, no. 13, pp. 4-15, 2021.

[24] Q. Zheng and M. Zwicker, "Learning to importance sample in primary sample space," in *Computer Graphics Forum*, 2019, vol. 38, no. 2, pp. 169-179: Wiley Online Library.