Dareen Alharthi

dareenharthi@gmail.com
(+1) 412-628-3603 or (+966) 580-669-560

EDUCATION

Master of Language Technologies , Carnegie Mellon University, School of Computer Science Language Technologies Institute, Pittsburgh, PA, USA August 2024 – August 2026

Bachelor of Computer Science, Imam Muhammad ibn Saud Islamic University, Riyadh, Saudi Arabia. First class honor, GPA **4.94/5**. October 2015 - July 2020

PUBLICATIONS

Tesselated Linear Model for Age Prediction from Voice ICASSP 2025
Authors: **Dareen Alharthi**, Mahsa Zamani, Bhiksha Raj, Rita Singh
Abstract: Voice biometric tasks, such as age estimation require modelling the often complex relationship between voice features and the biometric variable. While deep learning models can handle such complexity, they typically require large amounts of accurately labeled data to perform well. Such data are often scarce for biometric tasks such as voice-based age prediction. On the other hand, simpler models like linear regression can work with smaller datasets but often fail to generalize to the underlying non-linear patterns present in the data. In this paper we propose the Tessellated Linear Model (TLM), a piecewise linear approach that combines the simplicity of linear models with the capacity of non-linear functions. TLM tessellates the feature space into convex regions and fits a linear model within each region. We optimize the tessellation and the linear models using a hierarchical greedy partitioning. We evaluated TLM on the TIMIT dataset on the task of age prediction from voice, where it outperformed state-of-the-art deep learning models.

Evaluating Speech Synthesis by Training Recognizers on Synthetic Speech
SynData4GenAI Workshop 2024
Authors: **Dareen Alharthi**, Roshan Sharma, Hira Dhamyal, Soumi Maiti, Bhiksha Raj, Rita Singh
Abstract: Modern speech synthesis systems have improved significantly, with synthetic speech being indistinguishable from real speech. However, efficient and holistic evaluation of synthetic speech still remains a significant challenge. Human evaluation using Mean Opinion Score (MOS) is ideal, but inefficient due to high costs. Therefore, researchers have developed auxiliary automatic metrics like Word Error Rate (WER) to measure intelligibility. Prior works focus on evaluating synthetic speech based on pre-trained speech recognition models, however, this can be limiting since this approach primarily measures speech intelligibility. In this paper, we propose an evaluation technique involving the training of an ASR model on synthetic speech and assessing its performance on real speech. Our main assumption is that by training the ASR model on the synthetic speech, the WER on real speech reflects the similarity between distributions, a broader assessment of synthetic speech quality beyond intelligibility. Our proposed metric demonstrates a strong correlation with both MOS naturalness and MOS intelligibility when compared to SpeechLMScore and MOSNet on three recent Text-to-Speech (TTS) systems: MQTTS, StyleTTS, and YourTTS.

PAM: Prompting Audio-Language Models for Audio Quality Assessment
InterSpeech 2024

Authors: Soham Deshmukh, **Dareen Alharthi**, Benjamin Elizalde, Hannes Gamper, Mahmoud Al Ismail, Rita Singh, Bhiksha Raj, Huaming Wang
Abstract: Audio-Language Models (ALM) are pre-trained on a variety of audio-text pairs sourced mainly from the internet. Some of the text descriptions contain information about audio quality, the presence of artifacts or noise. Hence, ALM can take as input an audio recording and text prompts related to audio quality and assign a corresponding score. In this paper, we exploit this capability and introduce PAM, a no-reference metric for assessing audio quality for different audio generation tasks. Contrary to other "reference-free" metrics, PAM does not require computing embeddings with a pretrained model on a reference dataset nor training a task-specific model on an costly set of human listening scores. We extensively evaluate the reliability of PAM against established metrics and human listening scores on three tasks: text-to-audio (TTA), text-to-music generation (TTM), and text-to-speech (TTS). We perform multiple ablation studies with controlled distortions, in-the-wild setups, and prompt choices. Our evaluation shows that PAM correlates well with existing metrics and human listening scores. We break ground demonstrating the potential of ALM for assessing audio quality and how they can compute a general-purpose audio quality metric.

VERSA: A Versatile Evaluation Toolkit for Speech, Audio, and Music NAACL 2025 (submitted)
Authors: Jiatong Shi, Hyejin Shim, Jinchuan Tian, Siddhant Arora, Haibin Wu, Darius Petermann, Jia Qi Yip, You Zhang, Yuxun Tang, Wangyou Zhang, **Dareen Alharthi**, Yichen Huang, Koichi Saito, Jionghao Han, Yiwen Zhao, Chris Donahue, Shinji Watanabe.
Abstract: Neural codecs have become crucial to recent speech and audio generation research. In addition to signal compression capabilities, discrete codecs have also been found to enhance downstream training efficiency and compatibility with autoregressive language models. However, as extensive downstream applications are investigated, challenges have arisen in ensuring fair comparisons across diverse applications. To address these issues, we present a new open-source platform ESPnet-Codec, which is built on ESPnet and focuses on neural codec training and evaluation. ESPnet-Codec offers various recipes in audio, music, and speech for training and evaluation using several widely adopted codec models. Together with ESPnet-Codec, we present VERSA, a standalone evaluation toolkit, which provides a comprehensive evaluation of codec performance over 20 audio evaluation metrics. Notably, we demonstrate that ESPnet-Codec can be integrated into six ESPnet tasks, supporting diverse applications.

ESPnet-Codec: Comprehensive Training and Evaluation of Neural Codecs for Audio, Music, and Speech SLT 2024
Authors: Jiatong Shi, Jinchuan Tian, Yihan Wu, Jee-weon Jung, Jia Qi Yip, Yoshiki Masuyama, William Chen, Yuning Wu, Yuxun Tang, Massa Baali, **Dareen Alharthi**, Dong Zhang, Ruifan Deng, Tejes Srivastava, Haibin Wu, Alexander Liu, Bhiksha Raj, Qin Jin, Ruihua Song, Shinji Watanabe
Abstract: Neural codecs have become crucial to recent speech and audio generation research. In addition to signal compression capabilities, discrete codecs have also been found to enhance downstream training efficiency and compatibility with autoregressive language models. However, as extensive downstream applications are investigated, challenges have arisen in ensuring fair comparisons across diverse applications. To address these issues, we present a new open-source platform ESPnet-Codec, which is built on ESPnet and focuses on neural codec training and evaluation. ESPnet-Codec offers various recipes in audio, music, and speech for training and evaluation using several widely adopted codec models. Together with ESPnet-Codec,

we present VERSA, a standalone evaluation toolkit, which provides a comprehensive evaluation of codec performance over 20 audio evaluation metrics. Notably, we demonstrate that ESPnet-Codec can be integrated into six ESPnet tasks, supporting diverse applications.

LoFT: Local Proxy Fine-tuning For Improving Transferability Of Adversarial Attacks Against Large Language Model arXiv 2023
Authors: Muhammad A Shah, Roshan Sharma, Hira Dhamyal, Ankit Shah, **Dareen Alharthi**, Massa Baali, Hazim Bukhari, Joseph Konan, Soham Deshmukh, Bhiksha Raj, Rita Singh.
Abstract: It has been shown that Large Language Model (LLM) alignments can be circumvented by appending specially crafted attack suffixes with harmful queries to elicit harmful responses. To conduct attacks against private target models whose characterization is unknown, public models can be used as proxies to fashion the attack, with successful attacks being transferred from public proxies to private target models. The success rate of attack depends on how closely the proxy model approximates the private model. We hypothesize that for attacks to be transferrable, it is sufficient if the proxy can approximate the target model in the neighborhood of the harmful query. Therefore, in this paper, we propose *Local Fine-Tuning (LoFT)*, *i.e.*, fine-tuning proxy models on similar queries that lie in the lexico-semantic neighborhood of harmful queries to decrease the divergence between the proxy and target models. First, we demonstrate three approaches to prompt private target models to obtain similar queries given harmful queries. Next, we obtain data for local fine-tuning by eliciting responses from target models for the generated similar queries. Then, we optimize attack suffixes to generate attack prompts and evaluate the impact of our local fine-tuning on the attack's success rate. Experiments show that local fine-tuning of proxy models improves attack transferability and increases attack success rate by 39%, 7%, and 0.5% absolute on target models ChatGPT, GPT-4, and Claude respectively.

EXPERIENCE

Teaching Assistant for Introduction to Deep Learning Carnegie Mellon University
January 2024 – Present

- Lead recitation sessions to reinforce key concepts from lectures.
- Mentor students on course projects, providing guidance and technical support.
- Design and develop homework assignments to challenge and test students' understanding.
- Grade homework and provide constructive feedback.
- Hold office hours to address student questions and offer additional support.

Research Scholar Carnegie Mellon University August 2023 – July 2024

- Conducted research in speech synthesis and voice biometric estimation. Contributed to ongoing projects in the MLSP Lab with Prof. Bhiksha Raj, focusing on evaluating and benchmarking speech and voice representation models.

Deep Learning Team Lead at ISE October 2021 - August 2023

- Design ML systems and distribute tasks between team members.
- Research and implement the appropriate DL algorithms.
- Select appropriate datasets and representation.
- Analyze results for stability and validity.
- Create demos, documents, and present findings.

Deep Learning Engineer at ISE September 2020 - October 2021

- Enhance and build speech and computer vision systems on private real-world data.
- Implement models from scientific papers.
- Customize state-of-the-art systems on limited data.

Artificial Intelligence Engineer Intern at TAHAKOM June 2020 - September 2020

- Implement scripts to analyze satellite imagery.
- Research and present state-of-the-arts solutions for detecting vehicles from satellite imagery.
- Build and train object and change detection models for satellite imagery.

COURSES

General Assembly Data Science Immersive program June 2019 - September 2019
Machine Learning Modeling, Data Visualization, Data Mining, Python Programming, Corporate Client Engagement and Presentation Skills.

Coursera Deep Learning Specialization April 2020 - June 2020
CNN, RNN, LSTMs, Transformers, optimization strategies such as Dropout, Batch-Norm, and Xavier/He initialization.

Udacity Computer Vision Nanodegree February 2021 - April 2021
Feature extraction algorithms, and advanced computer vision scenarios such as object detection using deep learning models.

Coursera TensorFlow Advanced Techniques Specialization April 2021 - May 2021
Functional API, non-sequential models, Style Transfer, AE, VAEs, and GANs.