Audio Super-Resolution to Unlock the World's Languages

Wesley Maa, Palo Alto Senior High School

MIT Think Proposal

1.0 ABSTRACT

Approximately 6,500 languages are spoken around the world today. Yet, most speech recognition platforms using natural language processing (NLP) have only focused on seven key languages: English, Chinese, Urdu, Farsi, Arabic, French, and Spanish. As a result, 99.9% of the world's languages are increasingly being marginalized as NLP platforms become more entrenched in our daily lives. NLP research has historically focused on the above seven languages because high-quality voice datasets for the rest of the world's languages are generally more difficult to obtain. Instead, they are often stored in low-bitrate formats and require significant preprocessing before they can be used for modern NLP. The goal of my project is to develop a convolutional neural network (CNN) pipeline to solve an audio generation problem called bandwidth extension (aka "super-resolution"). Super-resolution transforms low-bitrate audio data into high-bitrate audio data that is then suitable for NLP. This project will be the first time a CNN is trained on a real-world voice dataset collected from phone calls into a customer-service call center in Southeast Asia, including multiple languages with multiple simultaneous speakers. By adapting CNNs to solve the problem of audio super-resolution, the long-term goal of this research is to make NLP accessible to all of the world's spoken languages.

2.0 MOTIVATION AND APPROACH

NLP plays a vital role in our everyday lives, from more natural interfaces to computing, to chatbots and language translation. However, the development of NLP has been far from equitable, as the majority of research has centered around only a handful of languages, particularly English. The development of modern NLP pipelines has been a challenge for the vast majority of the world's less popular languages because voice datasets are scarce and often require a significant amount of augmentation work that can be prohibitively expensive and time-consuming.

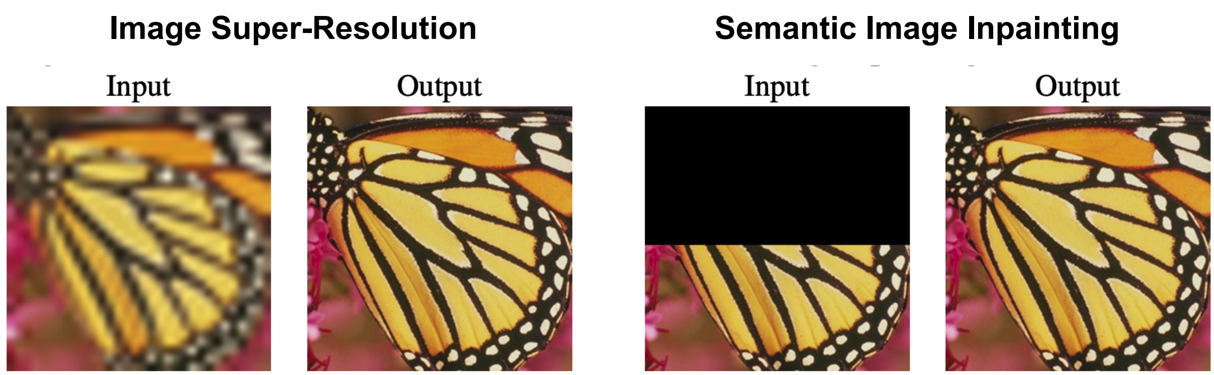
One commonly used NLP pipeline is Google's Bidirectional Encoder Representations from Transformers (BERT), which reports support for the top 100 languages in the world. However, the actual training of the system has only been performed on less than 20 languages because of the lack of high-quality training datasets (Romano, 2020). Other common NLP systems, including the Natural Language Toolkit (NLTK), have similar limitations.

Research in cognitive science has shown that language profoundly influences how we think. As a result, when NLP is limited to only a handful of languages, we implicitly program those selected languages' societal norms and biases into our digital platforms. By making NLP available to all of the world's spoken languages, not only are we democratizing access to all languages, but we also help to reduce the implicit biases that creep into our digital platforms.

In developing countries, customer call center recordings made over cellular or telephony networks are typically one of the most broadly available voice datasets. However, these datasets are unsuitable for training CNNs because of their legacy encoding standards. The vast majority of voice communications systems are based on traditional telephony standards developed in the 1950s, which limit the information bandwidth for voice communications to 300-3,400Hz. However, everyday conversational speech typically ranges from 0-8,000Hz, while the human ear can hear frequencies up to 20kHz. As a result, the audio codecs used in our phone networks encode voice calls at bitrates between 4-8kbps and capture less than 25% of the information generated by the human voice. This is why telephone conversations are less clear and more muffled than face-to-face conversations. To address this issue, telecommunication networks are increasingly upgrading their infrastructure to support high definition (HD) stereo-audio voice, which encodes at 12-40kbps using codecs like Adaptive Multi-Rate Wideband (AMR-WB). However, most mobile handsets, particularly those most often used in developing countries, don’t yet support HD voice codecs. Because the vast majority of voice calls remain encoded in low 4-8kbps bitrate data and contain a significant amount of background noise, these datasets have not been used for developing NLP pipelines for the longtail of the world’s 6,500 languages. Audio super-resolution recreates HD voice in software without the need to upgrade mobile phones or network infrastructure.

To construct an audio CNN, I’ve taken inspiration from the field of image analysis, where bandwidth extension is the process of constructing high-resolution images from low-resolution data. Before the availability of neural networks, this was accomplished through specific image processing algorithms like edge and focal detection. However, with the advent of CNNs, image bandwidth extension techniques have been developed that do not require domain-specific knowledge. Figure 1 provides an example of CNNs used for image super-resolution (Dong et al., 2016) (Kim et al., 2016) (Lai et al., 2017) and semantic image inpainting to predict masked regions in an image (Pathak et al., 2016) (Yeh et al., 2017).

Figure 1: Example of Image Bandwidth Expansion (Lim et al., 2018)

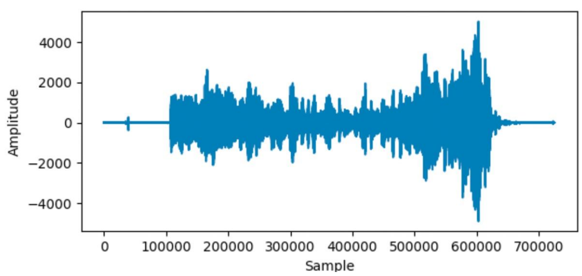


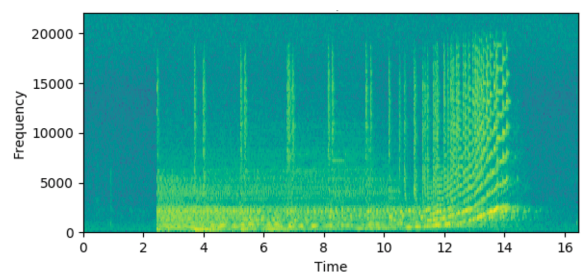
Similarly, in the field of audio analysis, bandwidth extension historically depended on traditional digital signal processing techniques for upsampling or feature extraction. This required a significant amount of domain-specific expertise and was difficult to generalize. Using CNNs, these traditional audio processing techniques are no longer needed. Instead, this project will first convert audio data into spectrogram images and then process these spectrograms through a standard image CNN pipeline developed for image analysis to achieve super-resolution. The resulting spectrograms are then converted back to audio data.

A spectrogram of an audio signal plots the frequency distribution of an audio signal (y-axis) over time (x-axis) and is an elegant way to capture audio data as an image. Different colors indicate the amplitude or strength of each frequency. Each vertical “slice” of a spectrogram is essentially the frequency spectrum at an instant in time. In Figure 2, the time-domain representation of an audio signal provides a sense of loudness over time but very little information about its frequency content. For comparison, the spectrogram of an audio file displays its frequency strength over time.

Figure 2: Decomposition of Audio Data (Doshi & Rosewell, 2021)

Time Domain **Frequency Domain (Spectrogram)**

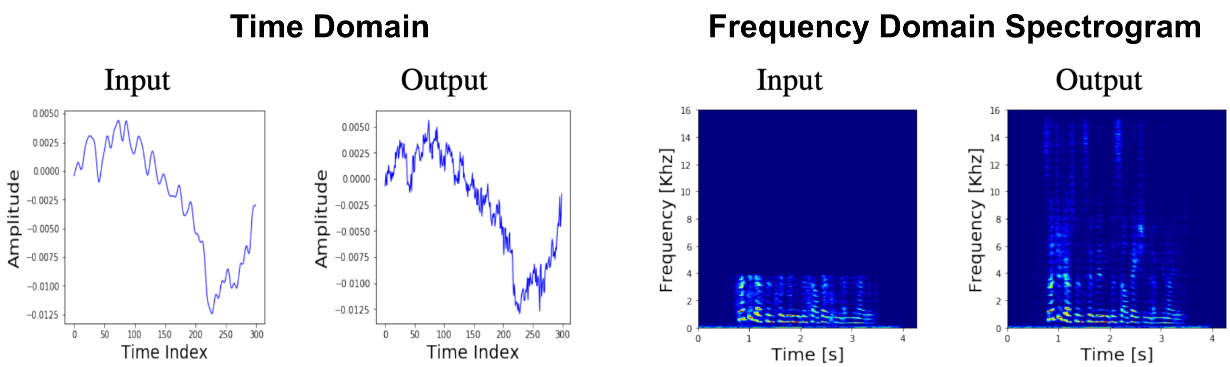




3.0 PROJECT LOGISTICS AND ORGANIZATION

In this project, I focus on a specific audio generation problem called bandwidth extension, in which the objective is to construct high-bitrate voice audio from a low-bitrate voice recording. To achieve this, I will apply a CNN pipeline developed for image super-resolution to voice-audio spectrograms to increase its sampling rate and frequency content. The CNN will be trained on pairs of low and high-bitrate voice-audio samples, where the low-bitrate input will be a telephone voice recording encoded at 4kbps, while the high-bitrate output will be a 4x upsampled reconstruction of the original voice recording at 16kbps. Figure 3 illustrates a comparison between a low-bitrate audio input and a high-bitrate audio output from existing CNNs for audio super-resolution (e.g., AudioUNet, AudioEDSR, and AudioUNetGAN). While these CNNs have been available for general audio recordings, this project will be focused on the specific application of training CNNs on voice recordings for NLP.

Figure 3: Comparison of low-bitrate audio inputs versus high-bitrate audio outputs (Lim et al., 2018)



For the training dataset, I will be using recorded voice data collected from the customer service call center at GRAB, a ride-sharing company based in Singapore that operates across eight different countries in Southeast Asia. This dataset is ideal for this project because Southeast Asia is home to over 1,000 distinct native languages. In fact, more than 800 languages are spoken in just the single country of Indonesia. This dataset contains over 364k hours of customer support calls across multiple languages in Southeast Asia and provides an extensive library of low-bitrate voice recordings. All voice calls are recorded at 4kbps, with a smaller subset recorded at 8-16kbps for use in the training process. This project will focus on assessing upsampling ratios of 2-4x to 8-16kbps.

This project will also assess AudioUNET on both the frequency and time domains of low-bitrate voice recordings to compare the validation accuracy of the two approaches because the two domains address very different problems. Audio super-resolution in the time domain is analogous to image super-resolution, which maps audio data from low-bitrate to high-bitrate. In contrast, super-resolution in the frequency domain is analogous to semantic image inpainting, i.e., given a particular set of low-frequency components in a spectrogram, reconstruct the high-frequency components. To exploit the best of both worlds, I plan to model and compare audio super-resolution in both the time and frequency domains.

To evaluate the quality of the output voice audio, I will use two different metrics, including signal-to-noise ratio (SNR) and log-spectral distance (LSD). The SNR function compares the CNN output against a high-resolution reference dataset and calculates the proportion of the output voice that is useful information in decibels. While SNR is simple to implement, there may be potential drawbacks, particularly when low-bitrate signals have significant background noise. To address this, I will also compare model outputs using LSD, which computes the amount of distortion present in the frequency spectra of a signal.

Over the past 12 months, I've successfully implemented a CNN pipeline for image semantic segmentation as part of an internship with the AI Lab at the National University of Singapore. In this project, I utilized a UNet architecture to train a supervised multi-class CNN. All pipelines were implemented in PyTorch. The NLP project proposed in this paper builds upon my prior work and extends the pipeline to perform super-resolution for voice audio. To prepare for this project, I've read several research papers (cited in the reference) to understand state-of-the-art CNNs for image super-resolution. Key milestones and their timelines for the remaining components of this project include:

Date	Milestone
February 29, 2022	Implement AudioUNET on the VCTK dataset (Dong et al.), a commonly used reference, to create a starting benchmark
March 31, 2022	Train CNN on GRAB datasets using a time-domain model and assess SNR and LSD metrics
April 30, 2022	Train CNN on GRAB datasets using a frequency-domain model and assess SNR and LSD metrics

Three key risks to this project have been identified, including:

1. Training artifacts from real-world datasets. This will be the first time a CNN is trained on a real-world voice dataset collected from customer-service call center recordings in Southeast Asia. Because of the poor quality of cellular networks and mobile handsets in Southeast Asia, this dataset contains significant noise and audio artifacts that may impact the accuracy of the CNN pipeline. To mitigate this risk, I may implement pre-processing filters to cleanse the dataset prior to use.
2. Combining time domain and frequency domain outputs. Time-domain and frequency-domain analysis address different aspects of bandwidth expansion (i.e., super-resolution vs. semantic inpainting). Improving model accuracy will likely require a combination of both models, and I continue to research an appropriate algorithm to merge the two approaches.
3. LSD and SNR metrics. LSD and SNR metrics may result in high model scores but still produce high-bitrate voice audio that sounds artificial. To address this limitation, this project will likely also evaluate model outputs using a mean

opinion score (MOS), whereby a group of people is asked to rate the overall quality of the voice outputs on a standardized scale. As a subjective metric, MOS measures how well a generative voice sounds, rather than just the accuracy.

My goal in applying to the MIT THINK program is to find mentors who can help advise on addressing the above key risks, particularly on 1) domain-specific audio pre-processing techniques, 2) the best algorithms to merge time-domain and frequency-domain model outputs, and 3) best practices for model and dataset tuning for CNNs. All three of these areas are skills that I will need to learn to complete this project.

Because this project is purely computational, I expect to complete all three milestones on a desktop PC. However, once this project transitions to training on GRAB datasets, the project may migrate to a cloud service using Amazon EC2 P3 instances to run the underlying CNN workloads. Funding for this project will go towards purchasing an Nvidia RTX 3080 GPU graphics card and potentially for EC2 credits.

Table 1. Project Budget

Item	On-Demand Price/hour	Available Hours in Budget	Total Cost	Link
Nvidia RTX 3080 GPU graphics card	NA	NA	$699.99	https://www.bestbuy.com/site/nvidia-geforce-rtx-3080-10gb-gddr6x-pci-express-4-0-graphics-card-titanium-and-black/6429440.p?skuId=6429440
EC2 P3	$3.06 / hour	98 hours	$300	https://aws.amazon.com/ec2/instance-types/p3/
Total			$1,000	

4.0 PERSONAL INTEREST

I've had an opportunity to live overseas for over 13 years as my family moved across Asia. During this time, I studied Bahasa for more than two years in Jakarta, Indonesia, and I studied Japanese for eight years in Tokyo, Japan. I also had an opportunity to immerse myself in the various dialects of Southeast Asia, including those in Singapore, Thailand, and Malaysia. Living abroad, I've developed a fascination for linguistics and its intersection with computer science. It's particularly eye-opening to see the large gap in access to NLP that exists between developed versus developing countries. Furthermore, this gap will only widen with time as modern machine learning pipelines essentially skip the developing world. Ultimately, my goal is to help shrink this gap and make NLP accessible to all of the world's spoken languages.

Works Cited

Dong, C., Change Loy, C., He, K., & Tang, X. (2016). Image super-resolution using deep convolutional networks. *IEEE Trans. Pattern Anal. Mach. Intell.*, *38*(2), 295–307.

Doshi, K., & Rosewell, J. (2021). *Audio Deep Learning Made Simple (Part 1): State-of-the-Art Techniques*. Towards Data Science. Retrieved December 20, 2021, from https://towardsdatascience.com/audio-deep-learning-made-simple-part-1-state-of-the-art-techniques-da1d3dff2504

Kim, J., Kwon Lee, J., & Mu Lee, K. (2016). Accurate image super-resolution using very deep convolutional networks. *The IEEE Conference on Computer Vision and Pattern Recognition*, *CVPR Oral*.

Lai, W.-S., Huang, J.-B., Ahuja, N., & Yang, M.-H. (2017). Deep laplacian pyramid networks for fast and accurate super-resolution. *IEEE Conference on Computer Vision and Pattern Recognition*.

Lim, T. Y., Yeh, R., Xu, Y., Do, M., & Hasegawa-Johnson, M. (2018). Time-Frequency Networks for Audio Super-Resolution. *IEEE International Conference on Acoustics, Speech and Signal Processing*, 646-650. 10.1109/ICASSP.2018.8462049

Pathak, D., Krahenbuhl, P., Donahue, J., Darrell, T., & Efros, A. (2016). Context encoders: Feature learning by inpainting. *Computer Vision and Pattern Recognition*.

Romano, S. (2020, June 20). *Multilingual Transformers*. Multilingual Transformers. https://towardsdatascience.com/multilingual-transformers-ae917b36034d

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *CoRR*.

Yamagishi, J. (2012). *English multi-speaker corpus for cstr voice cloning toolkit*. http://homepages.inf.ed.ac.uk/jyamagis/page3/page58/page58.html

Yeh, R., Chen, C., Yian Lim, T., Alexander G, S., Hasegawa-Johnson, M., & Do, M. N. (2017). Semantic image inpainting with deep generative models. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.