**Summary of the Thesis**

**Student Name:** Rahul Shukla
**Branch:** Electronics & Communication Engineering
**Guide:** Dr. Rajarshi Mahapatra
**Institute:** IIIT Naya Raipur

# Enhancing Neural TTS for Low-Resource Languages with Masked Language Models

This thesis explores advancements in neural text-to-speech (TTS) systems, particularly focusing on enhancing their applicability to low-resource languages. The rapid evolution of deep learning techniques has significantly improved the quality and naturalness of synthetic speech generated by TTS systems. However, the reliance on large amounts of paired text and speech data poses challenges for languages with limited resources. This research addresses these challenges by proposing a novel framework that integrates masked language models (MLMs) into the training process of Tacotron-based TTS models. The methodology involves a two-stage adaptation scheme: firstly, fine-tuning an MLM on text-only data from the target language to capture its linguistic characteristics; secondly, integrating these language-specific features into the embedding layer of the TTS model.

## Background and Motivation

This thesis addresses the challenges and opportunities in advancing neural text-to-speech (TTS) systems, particularly focusing on their application in low-resource languages. While neural TTS systems have significantly improved speech synthesis quality, they often rely on extensive paired text and speech data, limiting their accessibility for languages with limited resources. This disparity underscores the need for innovative approaches that can adapt TTS technologies to operate effectively with minimal data while maintaining or enhancing speech quality. Motivated by the goal of democratizing access to high-quality TTS across all languages, this research explores a novel framework integrating masked language models (MLMs) into Tacotron-based TTS models. By leveraging MLMs' ability to capture linguistic nuances from text data, the proposed methodology aims to enhance TTS adaptability and performance in diverse linguistic contexts, contributing to more inclusive and accessible speech technology solutions.

## Methodology

The methodology involves several key steps:

1. **Data Collection and Preparation:** The study begins with gathering text-only data from the target low-resource language. This corpus serves as the foundation for training the masked language model (MLM), crucial for capturing linguistic nuances specific to the language.

2. **Masked Language Model Fine-tuning:** Utilizing the collected text data, a pre-trained MLM (e.g., BERT) undergoes fine-tuning. This process adapts the MLM to the linguistic characteristics of the target language, enhancing its ability to understand and represent local language features effectively.

3. **Integration with Tacotron-based TTS Model:** The fine-tuned MLM embeddings are integrated into the embedding layer of a Tacotron-based TTS model. This integration augments the TTS model's capacity to generate speech that accurately reflects the linguistic nuances and context of the low-resource language**.**
4. **Model Training and Evaluation:** The enhanced Tacotron-based TTS model is trained using available paired text-speech data from the target language. Training focuses on o the model parameters to improve speech quality and naturalness.

5. **Performance Evaluation**: The effectiveness of the proposed methodology is evaluated through comprehensive performance metrics, including training and test accuracies, speech quality assessments, and comparisons with existing TTS approaches. This evaluation validates the framework's ability to enhance TTS capabilities in low-resource languages

**Experimental Setup and Results**

The experimental setup involved collecting text data from a low-resource language and fine-tuning a masked language model (MLM) on this dataset. We utilized a pre-trained MLM architecture and adapted it to capture linguistic nuances specific to the target language. The fine-tuned MLM embeddings were integrated into the Tacotron-based TTS model's embedding layer. Training of the enhanced TTS model was conducted using available paired text-speech data, focusing on optimizing parameters for improved speech synthesis quality.

The integrated approach demonstrated promising results with the Tacotron-based TTS model achieving a training accuracy of 97% and a test accuracy of 96%. Subjective assessments of synthesized speech quality highlighted enhanced naturalness and clarity compared to baseline models. These findings underscore the effectiveness of incorporating MLMs to enhance TTS systems for low-resource languages, offering a scalable solution to improve speech synthesis capabilities across diverse linguistic contexts.

**Conclusion and Future Work**

This thesis has explored a novel approach to enhancing neural text-to-speech (TTS) systems for low-resource languages by integrating masked language models (MLMs) into Tacotron-based architectures. The methodology successfully adapted MLMs to capture linguistic nuances specific to the target language, significantly improving speech synthesis quality. Experimental results demonstrated robust performance with high training and test accuracies, affirming the efficacy of the proposed framework. The enhanced TTS models produced speech that was noted for its naturalness and clarity, marking a substantial advancement in making TTS technology more accessible and effective across diverse linguistic contexts.

Moving forward, several avenues for future research and development are identified. Firstly, expanding the dataset size and diversity could further enhance the adaptability and generalization of the proposed framework. Additionally, exploring advanced MLM architectures and optimization techniques could refine model performance and efficiency. Integration with real-time applications and user feedback mechanisms could also validate the practical utility and

user acceptance of the enhanced TTS systems. Furthermore, extending the framework to encompass multilingual and code-switching scenarios would address broader language diversity challenges. These future directions aim to continue advancing TTS technology, making it more inclusive and impactful for global linguistic diversity.