**LEVERAGING WAV2VEC2 FOR TWI AUDIO TRANSCRIPTION: A Step Towards Building a Large Language Model for Low-Resource Language.**

**Abstract.**

Twi, a widely spoken language in Ghana, remains underrepresented in computational linguistics due to its classification as a low-resource language. This paper outlines an approach to leverage the Wav2Vec2-Large model for transcribing Twi audio and video content available on the internet. By generating large-scale textual data from abundant audio-visual resources, the project aims to contribute to the development of a robust Twi language model. This work serves as a foundation for addressing the lack of linguistic resources for Twi and sets the stage for creating large-scale natural language processing (NLP) applications for the language.

**1. Introduction.**

Languages with limited annotated data often face challenges in NLP model development. Twi, an Akan language spoken by millions in Ghana and beyond, lacks sufficient digital resources for computational analysis and model training. This project seeks to bridge this gap by utilizing advanced automatic speech recognition (ASR) technologies, specifically the Wav2Vec2-Large model, to transcribe Twi audio and video data freely available on the internet. The ultimate goal is to generate a substantial Twi text corpus to support the development of a large language model (LLM) for Twi.

**2. Background and Motivation.**

Twi's status as a low-resource language hampers efforts to build NLP tools like machine translation, voice assistants, and automated transcription services. Recent advances in self-supervised learning models, such as Wav2Vec2, have shown remarkable capabilities in ASR tasks for languages with limited data.

This project seeks to exploit these advancements to:

* Automatically transcribe Twi audio and video content.
* Build a sizable and high-quality Twi text dataset.
* Lay the groundwork for developing a Twi-specific large language model to support diverse NLP applications.

**3. Methodology.**

The project comprises the following key steps:

*3.1. Data Collection*

Twi audio and video content is sourced from publicly available repositories such as YouTube, podcasts, radio broadcasts and the internet (publicly available datasets). The focus is on ensuring diversity in accents, topics, and speech styles.

*3.2. Preprocessing*

Collected audio data is preprocessed to standardize formats and sampling rates. This includes resampling audio to 16 kHz, trimming silence, and normalizing volume levels.

*3.3. Model Training*

The Wav2Vec2-Large model, pretrained on multilingual data, is fine-tuned on the Twi dataset. Fine-tuning involves:

* Aligning phonetics and acoustic features unique to Twi.
* Utilizing available Twi text data for alignment and error correction.

*3.4. Transcription and Data Generation*

The fine-tuned model is used to transcribe large volumes of Twi audio and video data. Transcriptions are manually reviewed for quality assurance, ensuring high fidelity to the original speech.

*3.5. Dataset Creation*

Cleaned and validated transcriptions are compiled into a text corpus. This corpus is designed to capture the linguistic richness of Twi, including idiomatic expressions, tone, and syntax.

**4. Future Work**

The next phase involves leveraging the text corpus to train a large language model for Twi. Such a model would enable applications in machine translation, sentiment analysis, and conversational AI. Additionally, extending the dataset with domain-specific Twi content and integrating community feedback will enhance the model's robustness.

**5. Conclusion**

This project represents a critical step toward addressing the scarcity of Twi linguistic resources. By leveraging state-of-the-art ASR models and abundant online audio data, we aim to catalyze NLP research and application development for Twi, empowering speakers and preserving the language in the digital era.

Keywords: Twi, Wav2Vec2, Automatic Speech Recognition, Low-Resource Languages, Large Language Model, NLP.