GENERATING AUTO AUDIO TRANSLATION

SUGISIVAM.S[*1], VARUNAADARSH.M*[2] ,PRASANTH.A*[3].

[1.2.3.]UG scholar Department of Artificial Intelligence and Data Science Panimalar Engineering College, Chennai-600123.

ABSTRACT:

To translate every language into English is the project's fundamental goal. Our initiative can assist in translating other video languages into English. Our project entails 1).identifying the language, and 2). translating the language into text format. 3). transforming the text into an audio format 4). syncing the audio file. We employ a deep learning-based architecture called FuzzyGCP to identify languages. Hidden Markov Models (HMMs) for automated speech recognition (ASR) are used for audio to text conversion. We employ the Encoder, Decoder, and Converter components of Deep Voice 3 architecture to convert text to audio format.

INTRODUCTION:

IDENTIFYING THE LANGUAGE:

Detection of languages from the data, the techniques often entails examining numerous linguistic aspects of text data to identify the language it was written in. Language identification uses a variety of methods and algorithms, but some common ones are as follows:

N-gram analysis: This technique includes segmenting text data into shorter sequences of letters or words (referred to as n-grams) and comparing their frequency across various languages using statistical techniques. A machine learning system may be able to determine the language of a given text based on the frequency of certain n-grams, which may be more prevalent in Spanish than English.

Character-level characteristics: Another strategy is to analyse particular characteristics of the characters in a text, such their frequency, distribution, or even their shape.For instance, certain n-grams may be more prevalent in Spanish than in English, allowing a machine learning algorithm to estimate the language of a given text based on the frequencies of particular n-grams.

Character-level characteristics: Another strategy is to analyse particular characteristics of the characters in a text, such as their frequency, distribution, or even their shape. For instance, certain languages may utilise a higher proportion of particular characters or have particular patterns of diacritics that can be used to identify the language.

Neural network-based techniques: Since the development of deep learning, neural network-based models have grown in popularity for language identification. These algorithms learn to identify features within the information that are intriguing by being trained on vast amounts of labelled data.

TRANSLATING LANGUAGE INTO AUDIO FORMAT:

For many natural language processing activities, including translation, the conversion of spoken language into text is a crucial step. Automated speech recognition (ASR), a popular technique for this task, converts spoken words into written text using machine learning algorithms.

In order to represent the connections between speech acoustic data and the relevant linguistic units (such as phonemes or words), ASR systems often employ Hidden Markov Models (HMMs) or deep neural networks. These models are trained using substantial corpora of spoken language linked with corresponding transcriptions, enabling them to discover the statistical patterns shared by various languages and dialects.

The text produced by an ASR system once it has converted a spoken language into written text can subsequently be utilised as input for other natural language processing tasks, such machine translation. ASR is a difficult assignment, though, and mistakes are frequently made, especially when the speaker has an unusual accent or there is a lot of background noise. As a result, it's crucial to thoroughly assess the output of ASR algorithms and incorporate human experts in the transcription process to guarantee correctness and clarity.

AUDIO SYNCHRONIZATION:

Using audio signal processing methods in conjunction with machine learning algorithms is one possible method for synchronising an audio file using machine learning. The basic steps are as follows:

Preprocessing: Make the audio data ready for machine learning techniques by transforming the raw audio signal into a useful representation. You could, for instance, use an audio file conversion programme to create a spectrogram, which is a graphic depiction of the frequency content of an audio file over time.

A machine learning model can be trained by extracting pertinent features from the audio data. Aspects of the audio like the pace, beat, or pitch could be included in these features.

Training a machine learning model: Use a set of labelled audio files that have already undergone manual synchronisation to train a machine learning model. This model can become adept at identifying synchronization-related patterns in the audio features.

In order to forecast the synchronisation offset for a brand-new, unsynchronized audio file, use the trained model for synchronisation. To do this, it could be necessary to examine the unsynchronized audio file's features and compare them to the patterns developed during training.

Finalise the synchronisation by modifying the offset and making sure that the synchronised audio is positioned in line with the intended reference point.

In conclusion, synchronising an audio file with machine learning can be difficult and time-consuming, and it may call for specialised equipment and knowledge of machine learning and signal processing.

AUTOMATIC AUDIO TRANSLATION REQUIREMENTS:

Auto audio translation is a specialised area of machine learning and natural language processing that focuses on automatically translating spoken language in audio files. Several requirements must be met in order to accomplish accurate and trustworthy automatic audio translation:

Accurate voice recognition is the primary need for automatic audio translation. This entails verbatim transcription of the audio file's spoken words into text. High accuracy can be attained in this activity by utilising sophisticated speech recognition algorithms, including deep neural networks.

Language Translation: After the spoken language has been recorded as text, the target language must be added. Advanced machine learning algorithms that can evaluate the text's structure and content and produce precise translations are necessary for this. For this objective, neural machine translation models are frequently used.

Audio Alignment: The text output from the voice recognition and translation algorithms needs to be in line with the original audio file in order to provide reliable translations. This calls for specialised algorithms that can compare the timestamps of the uttered words with the corresponding words in the text.

Training Data: To hone the speech recognition and translation algorithms, a lot of training data is required. Along with data to train the alignment algorithms, this also consists of audio files with transcriptions and translations in the target language.

Quality Control: Automatic audio translation systems also need to implement quality control procedures to guarantee that the translations are correct and trustworthy. It may also entail routinely checking the system's performance and making necessary adjustments, as well as reviewing and correcting the output by humans.

Overall, sophisticated machine learning algorithms, a significant amount of training data, and close attention to quality control procedures are needed to achieve accurate and dependable automatic audio translation.

FUZZYGCP:

Despite being a method used for geo-registration in remote sensing and image processing, FuzzyGCP is not directly relevant for determining languages. Natural language processing (NLP) can, however, make use of fuzzy logic and fuzzy matching approaches to identify languages and raise the precision of language classification.

Using fuzzy logic, the fuzzy matching technique compares text strings that could include errors or variations and determines how similar they are. An unknown text can be compared to a collection of reference texts in several languages using this technique, and the degree of membership for each language is determined by how similar the unknown text and the reference texts are. The likelihood that an unknown text is one of each language can be calculated using this degree of membership, and the text can then be categorised accordingly.

Utilising a fuzzy decision tree is yet another method of language detection that uses fuzzy logic. Fuzzy logic is used in this decision tree approach to deal with ambiguous or unclear input data. A set of labelled text data in many languages can be used to train the fuzzy decision tree, which can then be applied to new text data to categorise it according to how much of each language it contains.

Overall, fuzzy matching and fuzzy decision tree approaches can be utilised in NLP to increase language classification accuracy and handle uncertain or ambiguous input data, even though FuzzyGCP is not applicable for language identification.

Hidden Markov Models (HMM):

A common method employed by automatic speech recognition (ASR) systems is the use of hidden markov models (HMMs). HMMs offer a probabilistic framework for modelling these acoustic qualities and predicting the appropriate text output in ASR systems, which analyse the acoustic characteristics of speech to convert spoken language into text.

The acoustic characteristics of speech are represented by a series of observations or feature vectors, such as mel-frequency cepstral coefficients (MFCCs), in an HMM-based ASR system. The sequence of observations is represented as a Markov chain, with the state at each time step being a latent variable that is not directly visible. Each observation is associated with a specific moment in time.

A set of states, each having a probability distribution across the data, make up the statistical model known as the HMM. The transitions between the states can be modelled using a transition matrix and are also probabilistic. The HMM is trained using a sizable corpus of speech data that has been labelled, with the labels correlating to the output text. Using methods like the Baum-Welch algorithm or the Expectation-Maximization (EM) algorithm, the training phase entails estimating the model's parameters, including the probability distributions for each state and the transition matrix.

By performing inference on the most likely sequence of states given the observed sequence of feature vectors, the HMM can be used to recognise speech once it has been trained. This is commonly accomplished using the Viterbi algorithm, which identifies the HMM's most probable path that corresponds to the observed feature vector sequence. The text matching to the HMM's most probable path is then the output of the ASR system.

Deep Voice 3 architecture to convert text to audio format:

For the purpose of converting text to voice (TTS), which entails creating natural-sounding speech from inputted text, Deep Voice 3 is a neural network architecture. To produce voice waveforms, the Deep Voice 3 architecture combines convolutional and recurrent neural networks with a different attention mechanism.

The architecture is made up of various major parts, such as:

Encoder: The encoder transforms the input text into a series of high-level linguistic properties, such as prosodic features and phoneme embeddings.

Decoder: The decoder uses the linguistic feature sequence to produce a series of acoustic features, like mel-spectrograms, which describe the speech signal.

The duration predictor uses the input text to estimate how long each phoneme in the speech signal will last.

The post-net is a convolutional neural network that improves the output of the decoder to create a voice waveform of greater quality.

In order to enhance the alignment of the linguistic and acoustic properties, the Deep Voice 3 design additionally makes use of a variation on the attention mechanism known as the directed attention mechanism. In order to direct the decoder in producing the acoustic features, this attention mechanism integrates knowledge of the predicted duration of each phoneme and the matching mel-spectrogram frame.

The Deep Voice 3 architecture, in its entirety, is a strong and adaptable method of TTS conversion that can produce natural-sounding speech with a high level of accuracy and fidelity. Virtual assistants, audiobook narration, and language learning aids are just a few of the applications that have made use of the framework.

CONCLUTION:

In this article, we present a technique for extracting speech signals and detecting spoken languages. Hence, we have created a

a deep learning-based ensemble architecture, called FuzzyGCP. A novel method in this field combines the use of SSGAN with the creation of an ensemble architecture. The audio classification ma

One of the important features of this architecture is its ability to solve the problem of image classification by using spectrograms. According to the findings provided in Section, the heterogeneous ensemble approach—which combines a traditional DDMLP as a classifier for numeric characteristics with classifiers for image-based features called DCNN and SSGAN—proved to be extremely useful.

The variety of datasets we have investigated, which include both Indic and other languages, demonstrate the durability and adaptability of FuzzyGCP. Unlike to its bi-lingual and tri-lingual counterparts in the field of SLID, a multi-lingual classification strategy is usually a difficult assignment that has been pretty successfully completed in this case. Yet, obstacles like the interstate

Connections between Indic languages, the use of a common vocabulary and demographic factors have an impact on

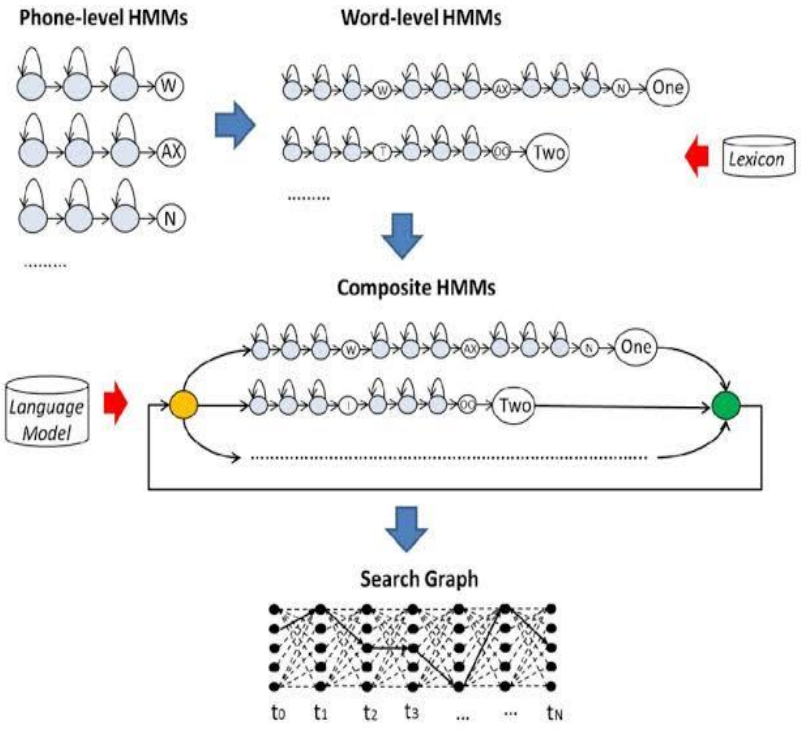
Having a deeper comprehension of languages is necessary. Improvement could be made in the future by utilising some using less computationally intensive structures and feature selection algorithms. implementation of sequential models like Gated Recurrent

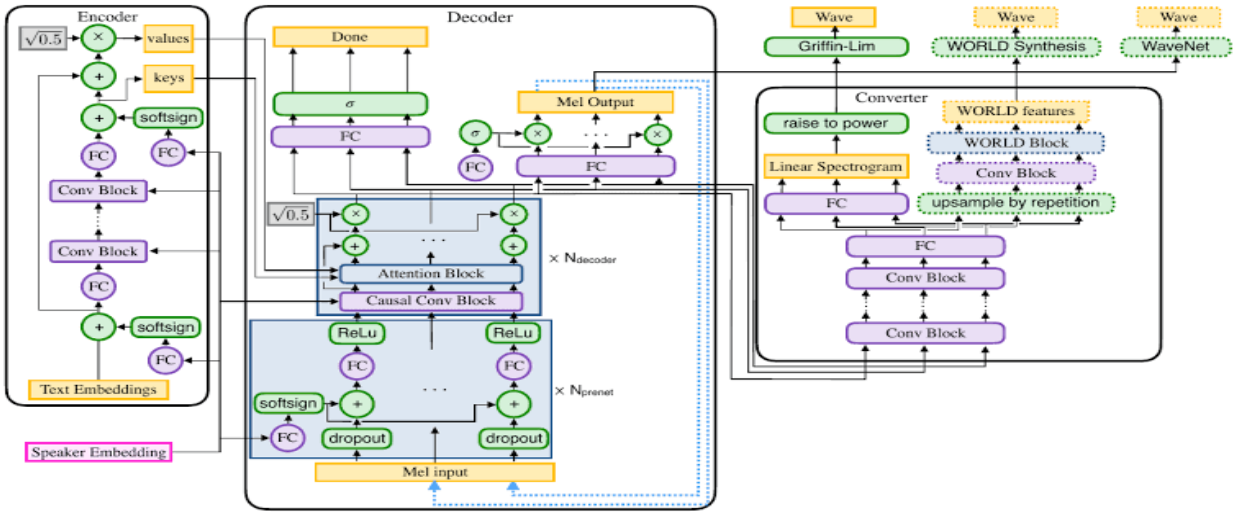
To create an ensemble, units (GRUs), LSTMs, etc. can be checked out.

On the datasets examined in this work, a thorough examination of the x-vector and i-vector based models with appropriate hyperparameter tuning is also a possibility. In addition, there are numerous other speech corpora available for study. A proper multilingual SLID system has the immediate advantage that it may be built on the output of the model for additional purposes, such as speaker profile development, automatic translation switching frameworks, ease of understanding in telemedicine purposes, etc.s

1. K. Drossos, S. Adavanne and T. Virtanen, "Automated audio captioning with recurrent neural networks," 2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, 2017, pp. 374-378, doi: 10.1109/WASPAA.2017.8170058.
2. A. D. McCarthy, L. Puzon and J. Pino, "SkinAugment: Auto-Encoding Speaker Conversions for Automatic Speech Translation," ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 2020, pp. 7924-7928, doi: 10.1109/ICASSP40776.2020.9053406.
3. L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," in Proceedings of the IEEE, vol. 77, no. 2, pp. 257-286, Feb. 1989, doi: 10.1109/5.18626.
4. K. Drossos, S. Adavanne and T. Virtanen, "Automated audio captioning with recurrent neural networks," 2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA), New Paltz, NY, USA, 2017, pp. 374-378, doi: 10.1109/WASPAA.2017.8170058.

DIAGRAM:




Encoder
Decoder
Converter
Wave
Griffin-Lim
WORLD Synthesis
WaveNet
values
keys
softsign
FC
Conv Block
Text Embeddings
Speaker Embedding
Done
Mel Output
Attention Block
Causal Conv Block
ReLu
dropout
Mel input
raise to power
Linear Spectrogram
WORLD features
WORLD Block
upsample by repetition