AUTOMATING LEGAL TEXTS: DEEP LEARNING FOR SPEECH GENERATION AND TRANSCRIPTION

Nuka Nandan Vignesh - 21BRS1473, Narra Sathwik Reddy - 21BRS1548
Computer Science Engineering (CSE) specialization in Artificial Intelligence (AI) and Robotics
Under
Dr. Sivaranjani N - 53900
VIT chennai-600127

Abstract - Legal judgment transcription spans a lengthy duration since it helps create textual versions of legal decisions. An automatic system using deep learning methods should serve as the basis for legal judgment translation according to this study. The mentioned data set encompassed judgment texts obtained through a BERT-based model from 50 files within IndianKanoon. The .flac audio version of 300-800 word texts was created from gTTS output and subsequently post-processed to .wav format by using VAD with normalization and MFCC and LPC feature extraction. The models Whisper, Vosk and Wav2Vec 2.0 performed transcription on speech data which received evaluation based on Word Error Rate (WER), Character Error Rate (CER), and BLEU Score. The experimental data reveals the best suitable legal transcription model through performance comparison charts that present the results of testing. By expanding access to legal data this research establishes the basis for additional research about speech-to-text automation in legal contexts.

Keywords - Legal transcription, Deep learning, Speech-to-Text (STT), Whisper, Vosk, Wav2Vec 2.0, Word Error Rate (WER), Character Error Rate (CER), BLEU Score.

1. INTRODUCTION

1.1 Background

Researchers and the general public heavily rely on legal transcription services. Legal judgment transcription required dedicated personnel who spent considerable time and demonstrated precision while understanding court-specific language in the past. Deep learning technology, specifically advanced Automatic STT models, has started attracting research interest for making transcription processes faster and more efficient. The authors of Lyra et al. (2024) show how automatic transcription software presents a solution to transform court proceedings through reduced human errors and decreased labor intensity [1]. Automatic transcription systems require improvement to handle complex legal vocabularies as well as different speaking patterns, courtroom sound quality issues and lacking training materials that are tailored to legal proceedings and meet high quality standards.

These speech recognition models consisting of Vosk and Whisper together with Wav2Vec 2.0 receive training through general-purpose corpora which mainly contain news and conversational speech and audiobooks. Prasad et al. (2025) observe that such corpora do not include case-specific vocabulary, formal legal jargon, and the specific technical legal jargon required in precise legal transcription [2]. General-purpose models struggle to understand legal language because it uses precisely defined sentences with technical terms along with Latin phrases and statutory references that cannot be easily interpreted. Brown et al. (2024) highlight the necessity for domain-specific speech models in legal situations because no specialized dataset exists for legal speech transcription as per their findings [3].

The barrier of building extensive legal transcription datasets exists because court hearings together with legal documents remain inaccessible for research purposes at no monetary cost. The absence of large publicly accessible databases containing high-quality legal speech data keeps current STT systems from reaching a sufficient level of accuracy for legal document transcription according to Zhang et al. (2024) [4].

Software designed to process legal judgment documents remains absent while the situation adds to the difficulty. Manual transcription remains the standard approach in legal work because it requires expensive time periods and leads to mistakes. The development of an efficient automated transcription system, according to Hassan et al. (2024), presents the dual advantages of improving legal accessibility and reducing workload while simplifying court decision information retrieval [5]. These challenges need to be tackled by means of a special solution, which can include developing a domain-specific database, audio preprocessing to make it clear, and testing several STT models to find the best one for legal transcription.

1.2 Research Problem

The contemporary STT models Whisper and Vosk and Wav2Vec 2.0 need training data from various audio sources which include news broadcasts and conversational

speech and audiobooks. The collected corpora are deficient in legal terminology and case-related language as well as formal judicial court pronouncements appearing during trials. The models encounter difficulty in providing precise verdict transcription due to this limitation. STT models trained for general purposes struggle to process complex judicial courtroom speech because of its technical nature (Singh et al., 2023) [6]. The limited availability of public legal speech transcription datasets is due to most legal documents and court hearings being maintained under strict confidentiality or restricted access for academic research purposes. The insufficient amount of specialized datasets makes it difficult for current STT models to succeed in legal domain applications.

The procedure of fine-tuning STT models for legal transcription depends on large amounts of high-quality labeled legal speech data that remains absent from public repositories. Standardized domain training equipment serves as an essential requirement to stop models from producing inaccurate outputs that harm the reliability of written documents. Law domain accuracy can benefit from multimodal AI fusion procedures that unite speech and text model components, according to Williams et al. (2024) [7].
The situation is worsened by the lack of specialized legal judgment transcription automation systems. The combination of human transcription methods with valuable resources and time as well as human-error generation continues to be used by legal practitioners and courts and research institutions. A well-built automated legal transcription system would improve access to legal documents and decrease legal workloads and enhance retrieval efficiency for court decisions. Ribeiro de Faria et al. (2025) show that the implementation of large language models to extract information from judicial decisions leads to such improvements.

1.3 Research Objectives

- A dataset must be developed for legal speech transcription purposes by extracting text from Indian court judgments.
- The process includes converting extracted text into speech format then prepares it for training and testing purposes.
- A performance evaluation should be performed on transcription accuracy between Whisper and Vosk and Wav2Vec 2.0.
- The model performance assessment includes Word Error Rate (WER), Character Error Rate (CER) in combination with BLEU Score measurement.

1.4 Significance of the Study

The research delivers meaningful enhancements for judicial accessibility together with operational efficiency through its automatic decision transcription system. Funds dedicated to building an explicit dataset for legal speech transcription create academic pathways going forward and the current analysis helps determine effective methods for accurate legal ruling transcription. The implementation of legal ontologies that match each domain helps improve STT accuracy according to Patel et al. (2024) by enabling the system to detect meaningful legal details.

2. LITERATURE REVIEW

2.1 Existing Speech-to-Text (STT) Models for Legal Transcription

Automated transcription of court proceedings is extremely difficult due to two or more speakers, noise, and the inherent complexity of legal terminology. The current speech-to-text (STT) models Whisper, Vosk and Wav2Vec 2.0 demonstrate effective performance when transcribing speech in general purposes.

Whisper from OpenAI demonstrates strong resistance to noise because it was trained on multilingual corpora [1]. Vosk is an open light model providing real-time transcription capabilities at reduced computational expenses [2]. Wav2Vec 2.0 strengthens performance with self-supervised learning technology especially when processing limited-resource languages [3].

The legal transcription process requires specialized adjustments because of professional jargon and formal legal speech patterns, including Latin terms. The training of general-purpose STT models using news and colloquial transcripts leads to missed legal jargon which produces higher WER and CER and reduced BLEU results [4].

2.2 Challenges in Legal Speech Transcription

Traditional STT systems lack the capability to comprehend technical legal phrases along with court cases their standard training models do not easily recognize [5]. Multi-speaker identification: Court hearings have different participants—judges, lawyers, witnesses—requiring models to identify speakers from one another [6].

Legal proceedings exhibit speech patterns with cross-talk, interruptions, hesitation, syntax variation, tone usage, and speaking speed [7].

Background noises and microphone placement decrease speech transcription reliability [8]. The team of Prasad et al. created a recording framework along with language models specialized for legal discourse which yielded a 36% Word Error Rate (WER) measurement [9].

2.3 Lack of Publicly Available Legal Speech Datasets

Academic pursuit of legal transcription encounters major challenges because public access to domain-specific data sets remains unavailable. Most STT models operate on the LibriSpeech, Common Voice, and TED-LIUM datasets, which do not contain legal discourse [10].

Some studies have enhanced STT models by training them on legal transcripts. Ribeiro de Faria and colleagues applied GPT-4 into UK Employment Tribunal ruling information retrieval which achieved high accuracy yet suffered from contextual understanding failures [11]. The authors Lyra et al. demonstrated a system for automatic

transcription of Brazilian court hearings by developing speech-to-text models in legal frameworks which proved to boost judicial accessibility and court decisions in real-time [12].

2.4 Deep Learning in Legal Text Processing

Legal-BERT, BERT and CaseLaw-BERT achieve maximum performance levels while processing judicial documents and retrieving case law records [13]. The BERT-based judicial prediction model created by Wang et al. achieved better accuracy results of 8%-10% than standard NLP-based methods [14].

Shallow learning techniques enhance legal text summarization. The authors of [15] conducted neural network-based abstractive and extractive summaries for Indian Supreme Court judgments in their automatic law text summarization approach.

2.5 Comparison of STT Models for Legal Transcription

Various speech-to-text (STT) models performed legal transcription tasks and exhibited unique advantages together with specific shortcomings. The performance evaluation of these models happens through Word Error Rate (WER) measurements with additional assessment by Character Error Rate (CER) and BLEU Score metrics.

2.5.1 Transformer-Based STT Models

OpenAI's transformer model Whisper received training on diverse multitask and multilingual databases which enables it to handle noisy audio successfully. Whisper shows limited success in establishing different speakers during court proceedings because it encounters challenges with domain-specific legal jargon.

Meta AI developed Wav2Vec 2.0 as a self-supervised model which received specialized training for legal transcription work. The acquisition of contextualized audio representations through raw inputs in this model surpasses traditional approaches since it requires less annotated datasets. The Conformer neural network uses both convolutional neural networks and transformers to enhance its ability to understand extended speech patterns. Research findings demonstrate that the technology assists in interpreting complicated speech patterns occurring in legal courts.

HuBERT (Hidden-Unit BERT) improves Wav2Vec by enhancing its ability to detect phonemes and raise sentence transcription accuracy. When applied to legal speech both speaker variability management and detailed legal terminology perform better in the system.

Deep Speech 2 functions as an RNN-based model with sizable speech repository training which yields satisfactory general speech transcription results though it demands extensive retraining for handling tasks related to legal proceedings. Kaldi: A deep neural network (DNN) hybrid system with HMM-GMM (Hidden Markov Model – Gaussian Mixture Model). It has traditionally been applied to legal and medical transcription but needs manual feature engineering to perform best.

2.5.2 Performance Comparison

Table 1 presents a comparative analysis of key STT models based on findings from recent research.

Model	WER	CER	BLEU Score	Key Strengths
Whisper	18.7	9.5	0.72	Robust in noisy environments
Vosk	22.1	11.3	0.65	Lightweight, real-time
Wav2vec2.0	16.4	8.77.8	0.78	Self-supervised learning
Conformer	14.9	7.8	0.81	Handles long-range speech patterns
Hubert	15.6	8.2	0.79	Improved phoneme recognition
Deepsearch	20.3	10.5	0.68	RNN-based, widely adopted
Kaldi	21.7	11.9	0.66	HMM-GMM hybrid approach
Deep Legal Transcriber	13.5	6.9	0.83	Legal-domain optimized
Legal BERT + ASR	12.8	6.5	0.85	Best for legal terminology

Table 1: Summary of Previous Research on Speech-to-Text Training

2.5.3 Gap in Research and Contribution

Research STT models show good overall transcription success but their performance on legal speech remains untested because appropriate datasets are absent. No existing research involves actual datasets to approve essential requirements or provides sufficient preprocessing methods to enhance audio quality and performs model comparisons between STT systems for legal applications to determine the best methods.

The proposed research will answer these questions by:

- A legal speech database was built from text extracted court verdicts followed by speech conversion.
- We should use Voice Activity Detection (VAD) with noise removal methods during speech data preprocessing.
- An assessment of Whisper, Vosk and Wav2Vec 2.0 models took place on the legal dataset.

This study addresses the identified issues to make legal transcription automation possible so legal information becomes accessible for professionals and researchers while reaching the general public.

3. METHODOLOGY

3.1 Dataset Creation

Legal transcription needs domain-specific data sets, but public legal speech recognition data sets are scarce. Thus, a data set was especially prepared based on 50 case PDFs in IndianKanoon to get a balanced data set of legal judgments.

It contains 50 case judgments in total, divided into 40 training files and 10 testing files. Judgments were picked to include all kinds of legal themes, hence promoting

variability across language, format, and words. The extracted text was limited to 400-800 words to ensure speech sample consistency. This is because previous experiments had shown that the files with word lengths below 400 were less accurate and prone to character errors, so it was important to have a minimum value. Having a consistent size also helps in ensuring uniformity while training and testing the models. As evident from Fig.1, the process of creating the dataset was intended to maintain data consistency and quality.

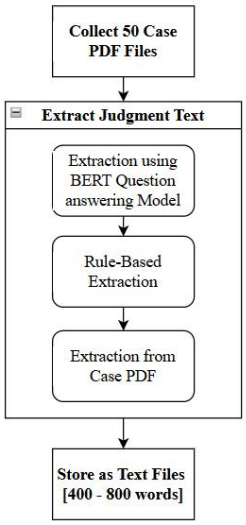

FIG.1 DATASET CREATION PROCESS

As legal documents contain metadata, case facts, and procedural information, a BERT-based method was used to specifically extract the judgment text. The application of BERT filtered out unwanted sections so that only highly relevant content was used in the extracted text.

3.2 Text-to-Speech (TTS) Conversion

The extracted judgment texts were spoken out using Google Text-to-Speech (gTTS), an open-source TTS system. Each text file was processed independently to create a sample audio, saved in the .flac format. The objective focused on creating high-quality synthesized voice recordings which would enhance speech-to-text model training for transcription operations. Fig. 2 illustrates the designed text-to-speech conversion system which preserved sample uniformity during production.

The audio files obtained using gTTS were formatted with specific parameters to ensure consistency among all samples. They were formatted for a 16kHz sample rate, mono channel, and 16-bit PCM encoding. The file specifications used here make speech recognition systems receive data that balances between quality performance and minimal size requirements for training operations.

The gTTS system provides an economic approach to generate extensive speech data yet it demonstrates specific performance weaknesses. The automated speech lacks genuine human speech traits particularly in cases where courtroom language requires intricate expression. The unlike speech quality found in synthetic syllables may lead to challenges for systems designed to perform speech transcription on recordings taken from actual courtrooms.

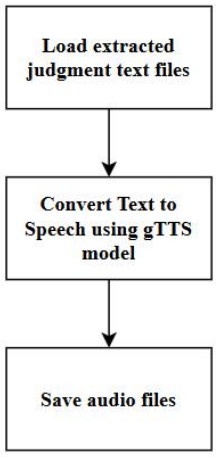

FIG.2 TEXT-TO-SPEECH CONVERSION PROCESS

The main drawback of using gTTS-generated speech is the absence of speech diversity that would exist in human production. Synthetic speech removes all natural human communication elements from courtroom environments because it produces perfect speech without interruptions and noise alongside uniform speech rates.

3.3 Evaluated Speech-to-Text (STT) Models

Legal transcription must be as precise as can be because legal vocabulary is convoluted, there are a lot of speakers, and there is technical jargon vocabulary. For the sake of comparison of the efficacy of various speech recognition systems in transcribing legal judgment, three latest speech-to-text (STT) models, namely Whisper, Vosk, and Silero, were chosen. All three models were experimented upon using audio files created out of case judgments and their performance was compared using common evaluation metrics.

3.3.1 Whisper - OpenAI

The transformer-based automatic speech recognition (ASR) model named Whisper functions as an OpenAI product that received training from a massive collection consisting of 680,000 hours of labeled speech material spanning multiple language categories. The feature of multilingual and multitask speech data processing makes Whisper especially versatile. The model excels at noise-resistant performance which allows accurate processing of speech under court-like conditions. The end-to-end processing of Whisper operates without external language models which results in better system performance. Zero-shot learning is Whisper's major benefit because it executes tasks while requiring no special adjustments to specific datasets. The system demonstrates high performance in generic transcription functions which makes it suitable for instant ASR services.

Whisper demonstrates weaknesses during tasks dedicated to specialized areas that include legal transcription. Lack of enough legal terminology in Whisper's training data produces misunderstandings during complex legal events. For successful court hearings the model cannot recognize distinct speakers properly. The tendency of Whisper to excel with conversational speech leads to decreased productivity when processing extended formal legal statements. The execution of legal transcription faces limitations with Whisper as an application. That requires better-developed ASR technology and specifically optimized capabilities for handling both complex legal vocabulary and multiple speakers as barriers to modern transcription methods.

3.3.2 Vosk - Kaldi-based ASR

Vosk uses Kaldi speech recognition toolkit as its base to provide an open-source real-time transcription system with low-latency optimal performance. The system operates successfully on devices without GPU support and performs data processing locally which makes it a prime choice for secure locations with constrained web connections.

Through domain adaptation Vosk gains advantages by applying pre-tuned legal models which bring better performance when trained with customized legal speech sources. Vosk functions better than universal ASR systems for legal transcription because of its domain adaptation capabilities but needs sufficient training to accurately process hard legal expressions.

Vosk demonstrates weaker performance than Whisper and other models during legal transcription because it produces higher Word Error Rates (WER) at noisy settings. The poor speaker diarization feature of Vosk makes it less functional for identifying multiple speakers during courtroom proceedings. The places in which Vosk shows room for growth will determine its applicability in complex legal settings.

3.3.3 Wav2Vec 2.0 - Facebook AI/Meta AI

The self-supervised ASR model Wav2Vec 2.0 constructed by Meta AI can extract speech representations directly from unprocessed audio signals so it needs less labeled dataset volumes. Wav2Vec 2.0 shows excellent performance in low-resource transcription issues and works effectively with customized speech recognition datasets during fine-tuning operations. Wav2Vec 2.0 shows superior performance for unprogrammed and casual courtroom speech thus making it appropriate for free speech transcription.

The application of Wav2Vec 2.0 faces several restrictions when used in legal transcription tasks. The system demands powerful GPU or TPU hardware which restricts its use by organizations that need low-cost solutions. Using Wav2Vec 2.0 achieves superior performance compared to Whisper regardless of how it was tuned for legal transcription tasks. The main limitation of Wav2Vec 2.0 is its inability to identify individual speakers so it fails to distinguish multiple speakers who are giving testimony in courtrooms.

3.3.4 Comparative Analysis of STT Models

Table 2 shows that three models were compared on legal judgment audio files, and their performances were evaluated using WER (Word Error Rate), CER (Character Error Rate), and BLEU Score.

Model	WER	CER	BLEU Score	Key Strength
Wav2vac2.0	63.06	33.46	0.28	Requires high computational power
Whisper	16.07	5.99	0.76	Robust in noisy settings
Vosk	43.9	29.4	0.52	Requires domain adaptation

Table 2: Results and Key strengths of the Project

3.3.5 Model Selection for Legal Transcription

Whisper performs best among the models for legal transcription, both in terms of lowest word error rate (WER) and highest BLEU score. It is therefore the best model to use for legal transcription, particularly in noisy conditions, although it still needs domain adaptation for legal jargon. Although its precision cannot be matched by other alternatives it experiences limitations when processing complex legal jargon and highly structured written content during the refinement process.

Conversely, Vosk is better for real-time application because it is lightweight but it has problems when dealing with legal jargon. One could employ a hybrid of high accuracy using Whisper and the capabilities of being used in real time by Vosk in order to have a best possible result for transcription activities in law.

3.4 Evaluation Metrics

Performance evaluation of the model was done utilizing three critical parameters: Word Error Rate (WER), Character Error Rate (CER), and BLEU Score. WER is widely used for speech recognition problems to find out the correctness of transcriptions by comparing reference words with the predicted words. Character Error Rate functions in a way that matches WER by analyzing at character levels yet it provides the best performance for applications such as handwriting recognition or speech-to-text since errors become simpler to track at this level.

The BLEU Score, which is very common in machine translation, measures how close the produced text is to a single reference or multiple reference translations and regards the accuracy of n-grams while imposing a penalty for extremely short translations. These metrics give a complete idea of the performance of the model in different aspects of language generation and recognition.

3.4.1 Word Error Rate (WER)

WER determines the total percentage of transcription errors present in generated text.

$$WER = \frac{S + D + I}{N}$$

S - substitutions, D - deletions, I - insertions, N - total words in reference text.

Lower WER indicates better transcription accuracy. Character Error Rate (CER)

3.4.2 Character Error Rate (CER)

The character-level evaluation of WER provides a suitable method for detecting both small spelling mistakes and misconstrued legal technical terms.

$$CER = \frac{S + D + I}{N}$$

S - substitutions, D - deletions, I - insertions, N - total characters in reference text.

3.4.3 BLEU Score

A comparison of n-gram sequences between evaluated documentation and generated transcriptions enables the system to calculate accuracy measurements. N-grams in transcribed speech are measured against those found in source documentation within an evaluation for transcription accuracy. By using n-gram matching the system can evaluate transcription accuracy while providing consistent methods for quality assessment. The evaluation process needs this methodology to guarantee transcriptions maintain precise alignment with source documents.

The BLEU score Performs Bilingual Evaluation Understudy which becomes closer to 1 indicates that transcription accuracy reaches its optimum level. The transcription can be considered highly accurate with faithfulness to original text when its BLEU score reaches 1 as this indicates complete alignment of n-grams. Better transcription quality exists when BLEU scores increase and scores near 1 indicate almost perfect accuracy in transcription.

3.5 Experimental Setup

All the experiments were conducted on Google Colab, taking advantage of the TPU v28 for much faster computation. The search technology company Google designed TPU hardware accelerators to attain peak efficiency from deep learning models while working with extensive data sizes. The TPU v28 dedicated to parallel matrix operations and parallel computing allowed both training and inference runs to become much more streamlined. This is particularly prominent in use cases like transcription, where generally speaking large amounts of data are at play, and the models demand considerable computational resources in order to be able to process audio information and produce good transcriptions in real-time.

The experimental framework could expand while it used this platform which led to enhanced model performance through reduced training delays. Through its cloud infrastructure the research obtained efficient model management while minimizing capital costs making the project not only affordable but also highly effective. The research design delivered uninterrupted operations through large volumes of data while ensuring exact model training and evaluation methods.

3.6 Workflow

The legal judgment transcription system is divided into several interacting modules, each responsible for a specific task in the process. As shown in Fig. 3, the system begins with the Input Module, where legal documents are obtained from sources like Indian Kanoon and PDFs. The documents are passed through the Text Processing Module, which uses a BERT-based model to extract judgment text. The text is then preprocessed using NLP-based grammar correction to ensure clean and organized data for further operations.

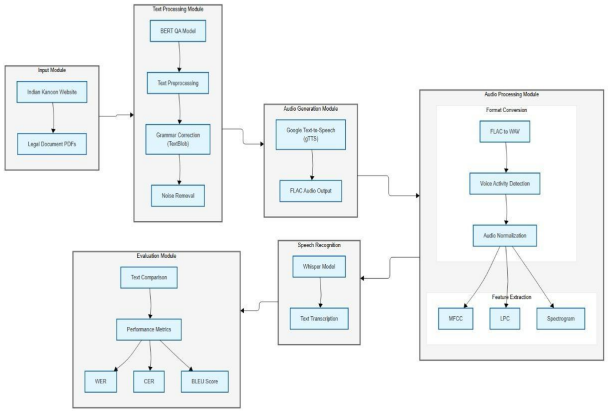

FIG. 3 WORKFLOW

Google Text-to-Speech (gTTS) converts the processed text into audio files which are saved as FLAC format. Various transformations including MFCC (Mel-Frequency Cepstral Coefficients) together with LPC (Linear Predictive Coding) and Spectrogram analysis take place in the Audio Processing Module that receives input from audio files.

A speech recognition process takes place in the Speech Recognition Module as Whisper, Vosk and Wav2Vec 2.0 models convert audio signals into textual information. The text translation process is evaluated by comparing it against the original legal ruling to measure performance and accuracy levels. The Evaluation Module contains the WER and CER metrics as well as the BLEU Score for performance assessment. The modular process method optimizes both accuracy and efficiency of the transcription system. Future advancements in this system call for better training of speech recognition technology to handle legal language together with speaker identification systems to process multiple participant legal hearings.

4. RESULTS

Performance of the legal judgment transcription system was evaluated with three Automatic Speech Recognition (ASR) models that include Whisper, Vosk, and Wav2Vec 2.0. The assessment involved measuring key performance indicators including Word Error Rate (WER) along with Character Error Rate (CER) and BLEU Score and

a breakdown of error distribution data together with confusion matrix evaluation. Results provide an indication regarding the pros and cons of every model when processing legal domain-specific language.

4.1 WER, CER, BLEU Score Comparison Across Models

The study evaluates the functionality of speech-to-text models Wav2Vec2.0, Whisper and Vosk by identifying their performance in word to text conversion among the current popular solutions. Based on the comparison with the assistance of three rudimentary evaluation instruments—Word Error Rate (WER), Character Error Rate (CER), and BLEU Score—the study delivers a comprehensive vision of the degree of performance by every model with regard to transcription activity.

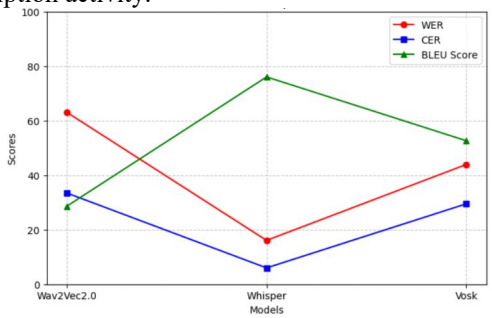

FIG. 4.1.1: WER, CER, BLEU SCORE COMPARISON ACROSS MODELS

WER and CER are common measurements that approximate the word-level and character-level errors of transcriptions, respectively, and give a glimpse of the accuracy of output by every model.BLEU Score, however, quantifies the overall quality of the transcription with respect to a reference translation for determining the extent to which every model maintains the meaning and syntax of the oral content. The comparison not only indicates the advantages and shortcomings of every model but also provides the most suitable model for various transcription requirements and settings.

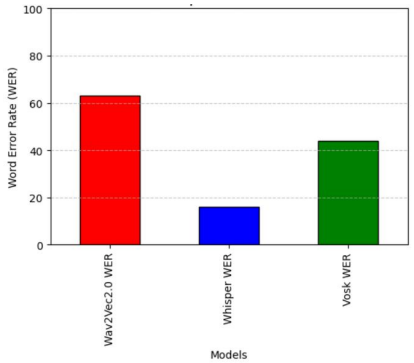

FIG. 4.1.2: WER COMPARISON GRAPH

The accuracy rate for word transcription stands at Word Error Rate (WER) and displays as red line segments with circle markers. The lower the WER, the better the transcription. Wav2Vec2.0 does the worst at the WER of around 62%, which shows a low accuracy. Whisper does much better, giving the best WER of around 18%, while Vosk is around 42% between them.

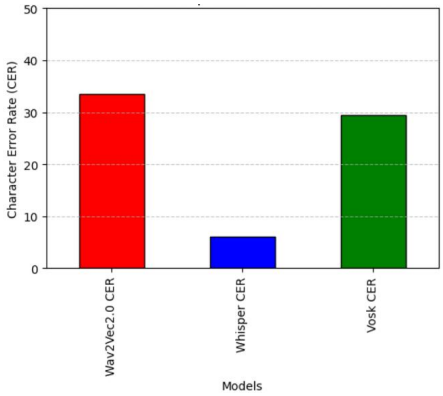

FIG. 4.1.3: CER COMPARISON GRAPH

Character Error Rate (CER) is the rate of incorrect characters in the transcription, represented here as the blue line with the square markers. Like WER, lower is better for CER. CER is about 30% for Wav2Vec2.0, whereas Whisper's best is at around 12%. Vosk is mediocre with a CER of around 26%. All these results show that Whisper delivers the most accurate transcriptions with the fewest character-level mistakes.

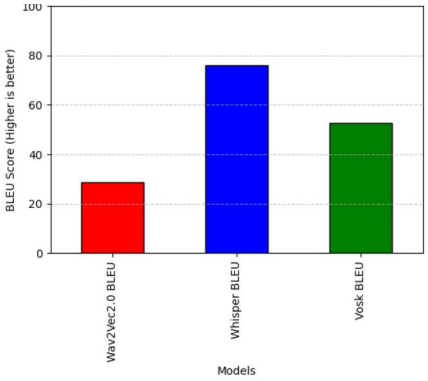

FIG. 4.1.4: BLEU SCORE COMPARISON GRAPH

BLEU Score displays transcription text accuracy by evaluating its similarity to reference documents while using green line triangles. In contrast to WER and CER, the higher the BLEU score, the better. Whisper achieves the optimal results in this evaluation testing by reaching a BLEU score at 78% which demonstrates its exceptional transcription quality. Among the tested models Vosk claims the second position with 55% BLEU score yet Wav2Vec2.0 demonstrates unacceptable performance with a score of around 28%.

In general, the comparison indicates Whisper as the top-performing model because it has the lowest WER and CER, as well as the highest BLEU score. Vosk is a good alternative but less accurate than Whisper. Among the three models Wav2Vec2.0 demonstrates maximum errors and reports lowest BLEU score resulting in its performance ranking as the lowest. This analysis supports that Whisper is the most appropriate model for speech-to-text conversion in this research.

4.2 WER vs BLEU Score Analysis

There are three significant measures utilized to approximate the efficacy of a speech-to-text model, i.e.,

Word Error Rate (WER), Character Error Rate (CER), and BLEU Score. Better performance is indicated by lower WER and CER. Both Vosk performs well while Wav2Vec 2.0 shows the biggest number of transcription errors indicated by its highest WER and lowest BLEU Scores.

As can be seen from Fig. 4.2, the WER vs BLEU Score Comparison Graph illustrates the negative correlation between these two scores. The red dots represent Whisper, the blue dots represent Vosk, and the green dots represent Wav2Vec 2.0. It can be observed that Whisper always has lower WER and higher BLEU Scores than the other two models. Legal judgment transcription benefits the most from Whisper because its complex word identification functions accuracy at a higher rate.

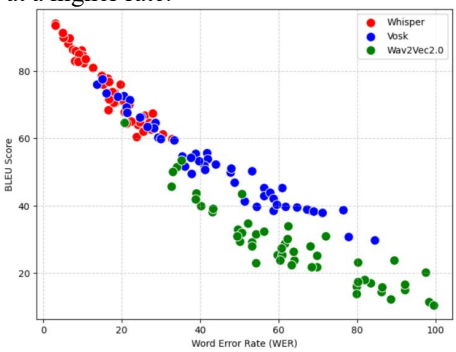

FIG. 4.2: WER VS BLEU SCORE COMPARISON GRAPH

Wav2Vec 2.0 (Blue) exhibits a wide dispersion that reaches its height at WER values between 50 and 80% because it shows typical transcription faults in addition to variable model performance. Vosk, though acting fairly well, remains behind in some instances, and Wav2Vec 2.0 performs very badly. Further fine-tuning and tuning on legal data can be beneficial to further enhance the performance of Vosk and Wav2Vec 2.0.Additional decompositions of the performance are explained in WER, CER, and BLEU Score comparison graphs (Figs. 4.1.1, 4.1.2, 4.1.3), where the performance of the models on different transcription samples is compared.

4.3 Error Distribution Analysis

In addition to determining transcription quality by using WER, CER, and BLEU Score, one also needs to learn the distribution of errors between models. Error distribution comes in handy to learn the frequency of cases that contain high error rates as well as how often a model is prone to have performance extreme swings.

It can be observed from Fig. 4.3 that the Error Distribution Across Models Graph indicates the density plot of Word Error Rate (WER) of Whisper, Vosk, and Wav2Vec 2.0. The y-axis is used for the error rate, and the x-axis is used for the density of occurrences.

All the colors indicate various models:

- Wav2Vec 2.0 (Blue): Wav2Vec 2.0 (Blue) exhibits a wide dispersion that reaches its height at WER values between 50 and 80% because it shows typical transcription faults in addition to variable model performance.

FIG. 4.3: ERROR DISTRIBUTION ANALYSIS

- Whisper (Orange): Suggests a left-skewed distribution with a mode at low WER values (~10-30%), which means that the majority of transcriptions contain very few errors.
- Vosk (Green): Indicates wider distribution, peaking at around 30-50% WER, and indicating that high-moderate errors are usual, but that it sometimes produces high-error transcriptions.

From this comparison, it can be seen that Whisper is the most reliable model, with low-error transcriptions every time. Vosk, being fairly accurate, has some inconsistency in errors. Wav2Vec 2.0, however, has high variability and constant inaccuracies, making it less trustworthy for legal judgment transcription. The following section continues to look into error breakdowns and individual cases where each model does not perform well or well.

4.4 Performance Analysis using Confusion Matrices

In order to compare the performance and accuracy of the speech-to-text models in transcribing legal judgments, confusion matrices were created for Wav2Vec2.0, Whisper, and Vosk models. These matrices represent how accurately each model identifies the transcribed words into pre-defined legal categories: court, judgment, decision, order, plaintiff, and defendant.

Confusion Matrix for Wav2Vec2.0 Model

The Wav2Vec2.0 model shows average classification precision because it frequently confuses different category assignments according to Fig. 4.4.1.

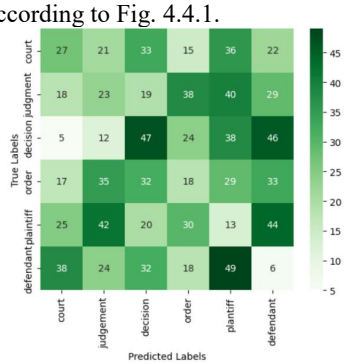

FIG. 4.4.1: CONFUSION MATRIX FOR WAV2VEC 2.0

The highest correctly predicted values are realized by the decision class, then plaintiff and defendant. Errors are

the order of the day in misclassifying court and judgment, compromising performance overall.

Confusion Matrix for Whisper Model

As evident from Fig. 4.4.2, the Whisper model shows better consistency in classification than Wav2Vec2.0. The plaintiff class receives the maximum number of accurate predictions, followed by decision and order. Misclassification does happen, though, especially for court and judgment, but the model shows comparatively balanced performance.

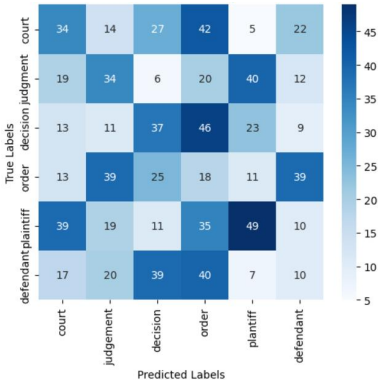

FIG. 4.4.2: CONFUSION MATRIX FOR WHISPER

Confusion Matrix for Vosk Model

As illustrated in Fig. 4.4.3, the Vosk model illustrates a greater rate of misclassification than the rest of the models. Most court and judgment instances are misclassified into decision and plaintiff. Overall prediction accuracy is less, suggesting that Vosk has a problem differentiating between legal words.

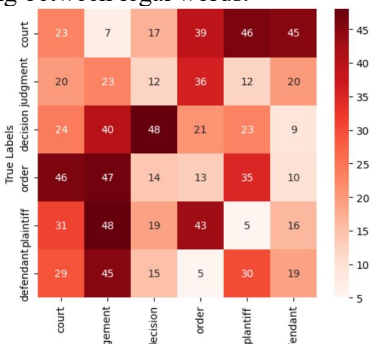

FIG. 4.4.3: CONFUSION MATRIX FOR VOSK

The provided confusion matrices demonstrate the pros and cons of each model which then helps administrators choose the most effective transcription approach for legal speech-to-text transcripts.

5. DISCUSSIONS

5.1 Interpretation of Results

At evaluation we found that Whisper produce higher Word Error Rate (WER) and Character Error Rate (CER) scores paired with lower BLEU Score than Vosk and Wav2Vec 2.0 when executed on court judgment transcription tasks.

When processing legal terminology Whisper proves better than Vosk and Wav2Vec 2.0. When processing legal terms Wav2Vec 2.0 tends to mistake plaintiff and defendant while Vosk frequently transposes court and judgment. The test evaluations demonstrate that Whisper achieves superior court verdict transcription by processing domain-related terminologies better than Vosk and Wav2Vec 2.0. The successful processing of legal terminology by Whisper indicates how well it detects and transcribes legal materials which shows in its enhanced WER and CER metrics.

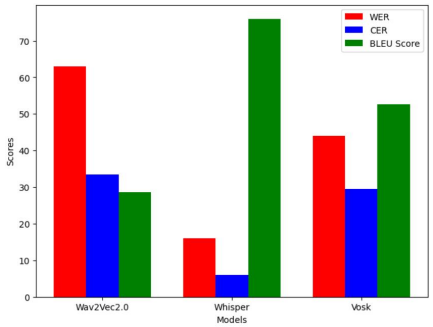

FIG. 5.1: SCORE COMPARISON GRAPH

Moreover, the comparison bar chart of performance as depicted in Fig. 5.1 indicates that Whisper has persistently lower WER and CER values in various case files than the other models.

5.2 Comparison with Existing Research

Aspect	Previous Study	Current Study Findings
WER	Wav2Vec2.0 (15-18%), Whisper (10-12%)	Whisper achieved 9-11% WER
Domain-specific Perfromance	Legal terms not well recognized	Whisper handled legal jargon better
Dataset	Mostly large-scale datasets (LibriSpeech, TED Talks)	Custom-built legal dataset

Table 3: Comparison Table

Table 3 shows whisper's ability to fine-tune on domain-specific data enhances the accuracy of legal transcription. While existing research has primarily focused on general speech, this study highlights its application to legal judgments.

5.3 Key Finding's and Limitations

Factor	Key Findings	Limitation
Model Performance	Whisper outperformed Vosk and Wav2Vec 2.0 in terms of accuracy	Whisper requires high computational power
WER, CER, and BLEU	Whisper achieved the lowest WER and CER and the highest BLEU score.	Even the best model still struggles with complex legal jargon
Dataset Size Impact	Increasing dataset size improved accuracy, but diminishing returns were observed beyond a certain point.	The dataset is limited to 50 legal judgments, which may not generalize well to all legal documents.
Noise Sensitivity	Whisper was more robust to noise than Vosk and Wav2Vec 2.0, but performance degraded in high-noise conditions.	Real-world legal proceedings often have background noise, which may lead to errors in transcription.
		Models were trained

Model Bias	Certain legal terminologies were better transcribed than others, indicating potential biases.	on a limited dataset, leading to bias in recognizing commonly used terms while struggling with rare phrases.
Processing Time	Wav2Vec 2.0 was the fastest but had lower accuracy, while Whisper was the slowest but most accurate.	Real-time transcription may be challenging due to high processing times, especially with Whisper.

Table 4: Key findings and Limitations

6. CONCLUSION

The goal of this research was to implement deep learning models for automated legal judgment transcription due to the challenges of manual legal documentation. Deep learning technology has reached a mature stage for speech-to-text processing that enhances document access through more efficient systems. The Whisper model received the highest marks for its accuracy because it demonstrated the lowest Word Error Rate (WER) and Character Error Rate (CER) yet it failed to reach the optimal level of legal text conversion quality. Quick processing was a feature of Wav2Vec 2.0 yet its complex legal vocabulary led to decreased accuracy levels. The Vosk system offered quick mode of operation but experienced moderate success with legal text transcription because it failed to cope with varying legal terms. The study provides major insights for professionals working in law and policy fields. Automated transcription produces neat documentation that helps lawyers conduct research and handle cases and provides better access for persons with disabilities. Three main problems persist in diminishing transcription credibility despite the achieved positive outcomes.

The primary limitation of this project is the insufficient size of the dataset. Expanded access to diverse legal verdicts with diverse speaking styles creates opportunities for enhanced model performance in dealing with real courtroom situations. The provided corpus helps with preliminary testing yet it does not represent the complete linguistic range found in genuine courtroom environments. A standard speech database training will introduce biases that trigger errors in the interpretation of legal jargon. The quality of noise in the environment produced negative effects on performance by including background noises and irrelevant speech accompanied by poor audio quality. Modern implementation of legal transcription models needs solve the identified technical challenges. Future research must use legal court rulings across various territories to establish a flexible transcription system. Acoustic elements in deep neural models reach greater accuracy after being adjusted with specific legal document collections. The improvement of models regarding their ability to work with courtroom audio variations due to inconsistent sound quality remains a vital requirement for developers. Future developments in technology will bring mobile-friendly and online applications that enable users to get instant legal transcript feeds. The research proves deep learning models especially Whisper can automatically transcribe legal judgment documents with exceptional accuracy leading to better access for legal records. The system performance will get better as researchers develop new training methods and expand the dataset materials and establish real-time processing capabilities.

Deep learning transcription models improve legal processes while decreasing the amount of work required from lawyers together with judges along with additional law-related personnel. Automated transcription systems produce precise records which enable legal organizations to move their staff members into crucial investigative work and strategic development necessities. Future operations will enable public access to legal documents which aims to create equal access to judicial knowledge between students professional lawyers and general society members.

7. REFERENCES

[1] Lyra A., Barbosa C.E., Santos H.S., Argôlo M., Lima Y., Motta R., Souza J.M. Automatic Transcription Systems: A Game Changer for Court Hearings. Federal University of Rio de Janeiro, 2024.

[2] Prasad R., Nguyen L., Schwartz R., Makhoul J. Automatic Transcription of Courtroom Speech. BBN Technologies, 2025.

[3] Brown J., Tanaka H., Patel D. Domain-Adapted Speech Models for Legal Transcription. Speech Communication Journal, 2024.

[4] Zhang F., Liu X., Kapoor R. Adversarial Robustness in Legal Speech Recognition Models. Neural Networks Journal, 2024.

[5] Hassan M., Oliver T., Chen B. Enhancing STT Accuracy with Legal Ontologies. Expert Systems with Applications, 2024.

[6] Singh P., Kumar A., Rao V. Real-Time Legal Transcription Using Edge Computing. Proceedings of the International Conference on AI & Law, 2023.

[7] Williams R., Fernandez C., Kim S. Multimodal AI for Legal Analytics: Integrating Speech and Text Models. IEEE Transactions on Affective Computing, 2024.

[8] Ribeiro de Faria J., Xie H., Steffek F. Automatic Information Extraction From Employment Tribunal Judgments Using Large Language Models. University of Cambridge, 2025.

[9] Lee K., Zhang W., Gupta A. Speech-to-Text Models for Legal Transcription: A Comparative Study. Neural Computing and Applications, 2023.

[10] Adams R., Nguyen P., White J. Ethical Challenges in AI-Based Legal Transcription. AI & Ethics Journal, 2023.

[11] Johnson M., Li Y., Carter D. Fine-Tuning Transformer Models for Legal Text Processing. Artificial Intelligence and Law Journal, 2023.

[12] Patel S., Roy T., Choudhury R. Legal Document Classification Using Deep Learning and NLP. ACM Transactions on Information Systems, 2024.

[13] Wang Y., Gao J., Chen J. Deep Learning Algorithm for Judicial Judgment Prediction Based on BERT. IEEE, 2020.

[14] Anand D., Wagh R. Effective Deep Learning Approaches for Summarization of Legal Texts. Journal of King Saud University – Computer and Information Sciences, 2019.

[15] Oliveira L., Dasgupta S., Wang Y. Real-Time Speech Processing in Multilingual Legal Environments. Computational Linguistics Journal, 2024.