



The
BRITISH
UNIVERSITY
IN EGYPT



London
South Bank
University

Faculty of Informatics and Computer Science
Artificial Intelligence

Project Title

Exploring Textless Speech-to-Speech: From Monolingual Correction to Cross-
Lingual Translation

By: Ahmed Imad

Supervised By

Professor. Andreas Pester

June 2025

TURNITIN REPORT

I understand the nature of plagiarism, and I am aware of the University's policy on this.

I certify that this report reports original work by me during my university project except for the following:

Ahmed Imad 226061 Thesis 3.docx

 British University in Egypt

Document Details

Submission ID

trn:old::11892:289350885

Submission Date

Jun 13, 2025, 11:16 PM GMT+3

Download Date

Jun 13, 2025, 11:18 PM GMT+3

File Name

Ahmed Imad 226061 Thesis 3.docx

File Size

70.1 KB

58 Pages

11,010 Words

61,160 Characters



Page 1 of 64 - Cover Page

Submission ID trn:old::11892:289350885







Page 2 of 64 - Integrity Overview

Submission ID trn:old::11892:289350885

8% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

Match Groups

-  **87 Not Cited or Quoted 8%**
Matches with neither in-text citation nor quotation marks
-  **11 Missing Quotations 0%**
Matches that are still very similar to source material
-  **0 Missing Citation 0%**
Matches that have quotation marks, but no in-text citation
-  **0 Cited and Quoted 0%**
Matches with in-text citation present, but no quotation marks

Top Sources

- 5%  Internet sources
- 3%  Publications
- 7%  Submitted works (Student Papers)

Integrity Flags

0 Integrity Flags for Review

Our system's algorithms look deeply at a document for any inconsistencies that would set it apart from a normal submission. If we notice something strange, we flag it for you to review.

A Flag is not necessarily an indicator of a problem. However, we'd recommend you focus your attention there for further review.

Signature

(Ahmed Imad)

Date 13/06/2025

Acknowledgments

I extend my deepest and most heartfelt gratitude to my supervisor, **Prof. Andreas Pester**. His invaluable guidance, constant encouragement, and insightful feedback were instrumental in navigating the complexities of this research. His mentorship not only shaped the direction of this project but also fostered my growth as a researcher, and for that, I am profoundly thankful.

I would also like to acknowledge the assistance of Google's Gemini, which served as a valuable tool for debugging minor coding errors and resolving complex software library conflicts, allowing me to focus on the core research challenges.

I also thank **Mr. Kyle from RunPod** for helping with RunPod's technical issues. Additionally, I am grateful to **Dr. Prabhjot** (WAYNE STATE University) for her assistance in providing technical assistance.

TABLE OF CONTENTS

1. Introduction	10
1.1 Overview	10
1.1.1 The Case for a Textless Approach	10
1.2 Problem Statement.....	10
1.3 Scope and Objectives.....	11
1.4 Motivation.....	11
2. Contributions	11
3. State-of-the-Art and Related Work.....	12
3.1 Traditional Cascaded Systems.....	12
3.2 The Shift to End-to-End Textless Models.....	12
3.3 Key Architectures in Textless S2ST.....	12
3.3.1 Direct Spectrogram Prediction.....	12
3.3.2 Unit-Based Approaches (S2UT).....	13
3.3.3 Advanced S2UT Models for Expressive Dubbing	13
3.4 Addressing Data Scarcity in S2ST	13
4. System Design and Architecture.....	13
4.1 Stage 1: Target Unit Generation (HuBERT + KMeans)	13
4.1.1 HuBERT: Self-Supervised Learning by Masked Prediction	14
4.1.2 K-Means Clustering for Unit Discretization.....	16
4.2 Stage 2: Speech to Unit Translation Model	16
4.2.1 Encoder Architecture	17
4.2.2 Decoder Architecture.....	17
4.3 Stage 3: Unit-to-Speech Synthesis (Vocoder)	17
5. Gantt Chart.....	18
6. Implementation	19
6.1 Hardware and Platform Selection: From Colab to RunPod	19
6.2 Environment Configuration & Reproducibility	19
6.3 Data Sourcing and Preparation.....	19
6.3.1 Dataset Selection and Sourcing	19
6.3.2 Storage and resource management.....	20
6.3.3 Workflow data processing	20
6.4 ASR Based Transcription	23

7. Evaluation Methodology & Metrics	24
7.1 Training-Based Metrics (loss vs. nll_loss).....	24
7.2 Text-Based Evaluation Metrics	24
8. Trials, Results, and Discussion.....	25
8.1 Initial Trials: Overcoming Architectural Incompatibilities	26
8.2 Trial 1: Monolingual Grammar Correction.....	27
8.2.1 Approach: A Novel Solution for Non-Paired Data.....	27
8.2.2 Evaluation: A Multistage Strategy.....	27
8.2.3 Adopting the Final Evaluation Metrics.....	36
8.3 Trial 2: Low-Resource Speech Translation (Arabic-to-English)	38
8.3.1 Setup and Rationale	38
8.3.2 Results and Discussion: A Successful Proof-of-Concept	38
8.4 Trial 3: High-Resource Speech Translation (German-to-English).....	39
8.4.1 A Macro Level Approach.....	39
8.4.2 Training Performance and Qualitative Analysis.....	40
8.4.3 Evaluation: A Comprehensive and Reliable Assessment	40
8.4.4 Discussion.....	43
8.5 Connecting Results & Comparative Analysis	43
8.5.1 Establishing a Baseline for a Novel Task (Grammar Correction)	44
8.5.2 Comparative Analysis against State-of-the-Art.....	44
9. Future Work	45
9.1 Exploring Advanced Model Architectures	45
9.1.1 State-Space Models (SSMs)	45
9.1.2 Unified Multimodal Models (e.g., Llama Omni).....	46
9.1.3 Diffusion and Flow Matching Models	46
9.2 Enhancing Expressiveness and Style Transfer.....	46
9.3 Data Efficiency and Augmentation	46
9.4 Refining Evaluation Methodologies.....	47
10. Conclusion and Recommendations.....	47
References	47
Appendix I	49
Appendix II	49
Appendix III	50

LIST OF FIGURES

Figure 1: Discrete Unit Generation (HuBERT + KMeans)	14
Figure 2: Discrete Unit Generation 2 (HuBERT + KMeans)	15
Figure 3: Hubert Cross Entropy Equation	15
Figure 4: Discrete Unit Generation 3 (HuBERT + KMeans)	16
Figure 5: Textless Speech-to-Speech Architecture	18
Figure 6: Gantt Chart.....	18
Figure 7: Manifest file for matching the data pairs (before deleting the suffix _en)	21
Figure 8: Source German WAVs and the English Corresponding Discrete units after matching.....	21
Figure 9: Source German WAVs and the English Corresponding Discrete units after matching (zoomed in)	22
Figure 10: Final .tsv file that includes source German speech with corresponding English discrete units before being passed to S2UT	22
Figure 11: Summarized Speech Workflow	23
Figure 12: Initial Trial - Bart Loss.....	26
Figure 13: Trail 1 , Grammar Correction BLEU, chrF2, TER, ROUGE, and METEOR scores on raw text (before normalization)	28
Figure 14: Trail 1 Grammar Correction - inference on raw text (before normalization)	29
Figure 15: Trail 1 Grammar Correction - inference 2 on raw text (before normalization)	29
Figure 16: Trail 1 , Grammar Correction BLEU, chrF2, TER, ROUGE, and METEOR scores on fully normalized text	30
Figure 17: Trail 1 Grammar Correction - inference 2 on fully normalized text	31
Figure 18: Trail 1 Grammar Correction - index matching check on fully normalized text	32
Figure 19: Trail 1 Grammar Correction - index matching check 2 on fully normalized text	33
Figure 20: Trail 1 Grammar Correction - index matching check 3 on fully normalized text	34
Figure 21: Trail 1 Grammar Correction - inference on partially normalized text (the success story)	35
Figure 22: Trail 1 Grammar Correction - BLEU, chrF2, TER, ROUGE, METEOR scores on partially normalized text (the success story)	36
Figure 23: Trail 1 Grammar Correction - WER score on partially normalized text (the success story)....	36
Figure 24: Trail 1 Grammar Correction - Final results Comparison Table.....	37
Figure 25: Trail 1 Grammar Correction - Loss and nll_loss Curve	38
Figure 26: Trail 2 Arabic to English Translation - Loss and nll_loss Curve	39
Figure 27: Trail 3 German to English Translation - Loss and nll_loss Curve.....	40
Figure 28: Trail 3 German to English Translation - BLEU, chrF, TER, ROUGE, and METEOR scores on raw data (before normalization).....	41
Figure 29: Trail 3 German to English Translation - WER score on raw data (before normalization).....	41
Figure 30: Trail 3 German to English Translation - BLEU, chrF, TER, ROUGE, and METEOR scores on fully normalized text	41
Figure 31: Trail 3 German to English Translation - WER score on fully normalized text	42
Figure 32: Trail 3 German to English Translation- Inference	42

Figure 33: : Trail 3 German to English Translation- Inference2	43
Figure 34: Trail 3 German To English - final results comparison table	43
Figure 35: Final results against State-of-the-Art - Comparison Table.....	44
Figure 36: Performance Vs SOTA	45
Figure 37: trail 1 Grammar Correction Output Speech (Snippets)	49
Figure 38: German to English Output Speech Snippets.....	49
Figure 39: project ownership and copyrights	50

Abstract

Textless Speech-to-Speech Translation (S2ST) is a future area of research that aims to translate speech between languages without passing through an intermediate text representation, usually by learning a source speech to discrete acoustic unit translation that is then synthesized by a vocoder. There are, however, major challenges to this approach, the first is the lack of large-scale parallel speech corpora, the second is that it is not trivial to judge a speech output with text based metrics, and ASR formatting decisions can cause poor scores even when the speech is of high quality. This project sought to deploy and comprehensively evaluate a textless Speech-to-Unit Translation (S2UT) pipeline on various domains to understand the strengths and weaknesses of such a system. The main achievement of this work was shown on a monolingual grammar correction task, where a textless methodology was designed to synthetically create a paired dataset, resulting in a high ROUGEL score of 85.10 (scale 1-100). This effectively proved the feasibility of the S2UT pipeline in complex monolingual speech transformation. The pipeline was subsequently evaluated on cross-lingual translation to understand its data scale dependency; an experiment on a low-resource Arabic-to-English task produced output that was clearly noisy when listened to, which supported the hypothesis that large-scale data is required. The model was then trained on a large-resource German-to-English dataset of approximately 128,000 examples, at which point the audio fluency considerably improved, along with the BLEU score of 28.21. Nevertheless, the model continued to make some wrong translations, and it can be concluded that the scale of data is a vital aspect, but to receive high-fidelity translation, this architecture might demand even more extensive datasets or additional architectural improvement.

1. Introduction

Human beings have an incredible, inherent capacity not only to comprehend spoken language but also to hear and make corrections to grammatical errors on a real-time basis. This mental phenomenon of listening to a wrong phrase and converting it in the mind into a right one is a pillar of communication and acquisition. To computers, though, the task of carrying out such transformations on audio waveforms is daunting. In the modern era of artificial intelligence, directly operating end-to-end sequence-to-sequence models, computers can perform speech-to-speech (S2S) tasks without an intermediate text pipeline.

The proposed project will realize, analyze and assess a textless S2S pipeline along the Speech-to-Unit Translation (S2UT) paradigm. The objective is to take a source audio waveform and produce a new audio waveform in a target language or a corrected form. A key contribution of this thesis is the novel adaptation of this conventional speech-to-speech translation framework. We re-use the S2UT pipeline, which was originally designed to solve tasks such as German-to-English translation, to solve a different problem of monolingual grammatical error correction in speech, thus showing its applicability beyond cross-lingual tasks. Such a system when successful can have huge applications in fields such as real time translation devices, language learning aids and communication assistive devices.

1.1 Overview

This project uses the textless Speech-to-Unit Translation (S2UT) model as the speech generation pipeline. This system consists of multiple deep learning modules that operate stages: a Speech Encoder encodes the source audio signal, a Transformer Decoder decodes it into a sequence of discrete acoustic units, and a Vocoder generates these units into the target audio waveform.

1.1.1 The Case for a Textless Approach

One of the early decisions that form the basis of this project is the S2S direct pipeline. Conventional systems approach this task in a cascaded manner (Speech-to-Text -> Translation/Correction -> Text-to-Speech), which allows the errors to accumulate at every step. While effective, this approach suffers from several drawbacks, including the compounding of errors from each stage, higher latency, and the complete loss of paralinguistic information such as emotion, tone, and speaker identity from the original source speech. The challenges of a text-based pipeline were made apparent even in the final evaluation stage of this project; we struggled in using the text just for the evaluation due to ASR errors and complex normalization issues. This raises a critical question: if text-based evaluation alone presents such significant challenges, what are the implications of designing the entire system around textual representations? The outcome could be highly suboptimal.

Thus, it was decided to go directly, without texts, and this is an innovation and a special paradigm in the sphere. It seeks to avoid the problems of text-based systems by instead learning a direct mapping between source speech and target speech acoustics. Chose to base a project on such a cutting-edge technique, considering the limited resources and the difficulty of obtaining appropriate large-scale datasets, was a big decision to make, which this thesis managed to pull off.

1.2 Problem Statement

Given a source audio utterance, produce a high-quality fluent speech utterance that is either a correct translation into a target language or a grammatically corrected version of the source language, without intermediate text representations in the pipeline core model.

1.3 Scope and Objectives

The project will focus on enhancing the quality of the produced speech by realizing an effective S2UT pipeline and scientifically studying its behavior on various tasks and data settings. The primary ones are:

To deploy an entire end-to-end textless S2UT pipeline with state-of-the-art components such as Fairseq, HuBERT and HiFi-GAN.

To explore the pipeline on the new task of monolingual grammar correction by creating a method of synthetic data generation.

To examine how the pipeline decorated with data scale through evaluating its performance on low-resource (Arabic-to-English) and high-resource (German-to-English) translation tasks.

To design and evaluate an effective ASR-based assessment system to find a solution to the problem of textless output.

Important to note, the initial scope of the project also envisioned the possibility of a dialogue-based, interactive continuation of a chat. This idea was however scoped out since a chat continuation feature is much more complicated and would require a large dataset of million records to make such an endeavor successful. It was decided to consider only single-utterance translation and correction, as, firstly, it is necessary to obtain a strong baseline model.

1.4 Motivation

Verbal communication is the most basic and the most natural way of human communication. As the world becomes more digitalized, speaking may be much more convenient than typing, and speech is certainly the definite future of human-computer interaction and worldwide communication. Yet the very nature of traditional translation systems involving an intermediate step based on textual representation inevitably loses much of the richness of such interaction. According to what has been written on the subject, critical paralinguistic information, which includes emotion, tone, rhythm, and distinctive speaker identity, is discarded when it is transcribed into a text form.

A novel idea that is expected to address these shortcomings is the textless Speech-to-Speech Translation (S2ST) pipeline. This methodology has the advantage of entraining the crucial, expressive aspects of speech that are not residual in text, as it learns to directly map between a source audio waveform and a target audio representation. And though it will be a difficult task, as this thesis will show, this approach opens the gates to a more natural and interesting future of cross-lingual communication.

Direct translation of speech is a big frontier of machine learning, as it breaks the existing text-based paradigms and tries to capture a far richer and more complex signal. The main reason to address this challenge is the seek of expressive translation. A conventional cascaded system is able to pass the literal meaning of words, whereas a real Speech-to-Speech Translation (S2ST) system attempts to translate the whole act of communication: the emotional state of the speaker, his/her intent, and identity, which are all encoded in the prosody and timbre of the voice . Through a textless direction, this piece intends to bring its grain of salt to a more natural, timely, and barrier-free communication in the future.

2. Contributions

to the project Some important contributions of the field of textless speech processing are offered by this project:

The main contribution of this work is that we have successfully applied and end-to-end validated a textless S2UT pipeline on multiple and difficult tasks. This has been done in the face of great challenges such as the uniqueness of the research topic, scarcity of resources in form of publicly available parallel speech corpora and the intensive computation requirement that is many times more than that of an ordinary project.

An entirely textless approach to synthetic data generation was designed. This direct manipulation of discrete acoustic units to introduce the simulated errors allowed us ***to construct a large-scale paired corpus of the grammar correction task with no use of text, and can be potentially generalized to other data-scarce speech transformation tasks.*** Leading to a novel approach, *there are no existing papers applied the grammar correction to textless pipeline.*

Through a controlled comparison of low-resource and high-resource training, we provided a clear, quantitative demonstration of the S2UT pipeline's dependency on data scale. We concluded that for complex cross-lingual tasks, datasets with over 100,000 samples are a critical prerequisite for achieving high-fidelity translation.

3. State-of-the-Art and Related Work

The chapter is an overview of the major developments and activity in the domain of Speech-to-Speech Translation (S2ST), which sets the scene of this project. We shall discuss the historical progression of the traditional systems to the current end-to-end textless methods, the major architecting paradigms, and latest tactics of beating data paucity.

3.1 Traditional Cascaded Systems

The traditional way to implement S2ST is a sequence of three autonomous models: Automatic Speech Recognition (ASR), Machine Translation (MT) and Text-to-Speech (TTS). This cascaded approach though modular has serious disadvantages. First, mistakes are cumulative; an ASR transcription mistake may result in a nonsensical translation, which will then be pronounced fluently by the TTS, resulting in a poor final product. Second, such a pipeline is text-centric by nature: all of the paralinguistic information present in the source speech, including emotion, tone, rhythm, and speaker identity, is discarded at the ASR stage. Lastly, the three models are sequential, which increases the latency and unsuitable in real-time applications.

3.2 The Shift to End-to-End Textless Models

To alleviate these drawbacks, the current study has trended to end-to-end S2ST models. Such models seek to directly learn the mapping between a source speech waveform and target speech representation (omitting the intermediate step of generating text). The textless method offers latency reduction, lessened error spreading, and expressive translations that are capable of maintaining vocal qualities. This modern paradigm is where the work in this thesis is placed.

3.3 Key Architectures in Textless S2ST

3.3.1 Direct Spectrogram Prediction

The most basic method is to learn a model which directly regresses the spectrogram of a source utterance to a target spectrogram. The first model in this direction was the Translatotron [1] by Google. It was the first to show that a direct method using attention-based sequence-to-sequence model could preserve the

voice of source speaker. Translatotron 2 [2] was an improvement over the original in every meaningful way, using a superior spectrogram synthesis algorithm and a higher-quality vocoder.

3.3.2 Unit-Based Approaches (S2UT)

The Speech-to-Unit Translation (S2UT) is the state-of-the-art paradigm chosen as the basis of this thesis. The method, promoted by a team of researchers at Meta AI and others, factors the task into two separate steps: translation and synthesis. The source speech is then translated into a sequence of discrete acoustic units (obtained with a self-supervised model such as HuBERT [5]) with a sequence-to-sequence model. This unit sequence is then transformed into a high-quality waveform using a separate vocoder, e.g. HiFi-GAN. The procedure makes the translation task easier and has been shown to scale well, with a 7-billion-parameter textless S2ST model having been trained [7].

In order to achieve faster inference speed as well, researchers have also proposed non-autoregressive (NAR) models such as TranSpeech [8], which parallel predicts all units with a new bilateral perturbation mechanism, showing a large reduction in latency.

3.3.3 Advanced S2UT Models for Expressive Dubbing

It is worth emphasizing that the tempo of innovation is quite fast, and extremely recent research has adapted the S2UT pipeline to certain and demanding tasks, such as dubbing. An example is the Dub-S2ST model [6] that incorporates a number of important innovations in order to maintain the timing and prosody of the source speech. It explicitly controls the duration of the output by using a discrete diffusion model and a Diffusion Transformer (DiT) decoder. More so, it employs a new unit-based speed adaptation algorithm and a Conditional Flow Matching (CFM) synthesizer to preserve the original speakers pace and speaker identity, which makes it so appropriate in Dubbing without notice .

3.4 Addressing Data Scarcity in S2ST

One of the main problems in all S2ST methods is the unavailability of large-scale parallel corpus of speech. A popular way of data augmentation is a text-based pipeline: translated text is converted to speech with a TTS system [9]. This is as opposed to the new approach developed in this thesis in the grammar correction task. We have preferred to keep our system completely textless, by operating directly on the discrete acoustic units of clean speech to introduce errors programmatically. The corrupted unit sequences resulted were subsequently vocoded to produce the source audio. This establishes the fact that at scale synthetic paired data for speech-based tasks can be generated without text dependency.

4. System Design and Architecture

In this chapter, the proposed solution is broken down in detail: an end-to-end pipeline of textless Speech-to-Unit Translation (S2UT). The architecture consists of a number of components which are state-of-the-art and each component is selected to tackle a particular aspect of the S2ST challenge. Its general architecture is strongly based on the effective textless translation systems proposed by the researchers of Meta AI and Google [5, 8], and the model has three major phases: Target Unit Generation, Speech-to-Unit Translation, and Unit-to-speech Synthesis.

4.1 Stage 1: Target Unit Generation (HuBERT + KMeans)

The initial step in the S2UT pipeline is the generation of a target vocabulary of discrete acoustic units. It is a type of representation learning in which the space of speech, which is high dimensional and continuous, is mapped to a discrete set of tokens which is finite. The method is based on a self-supervised speech model, HuBERT, and then K-Means clustering.

4.1.1 HuBERT: Self-Supervised Learning by Masked Prediction

The unit generation process is based on the HuBERT (Hidden-Unit BERT) model [10]. HuBERT is trained in large quantities of unlabeled audio data to learn speech representations that are rich. It is designed on the Transformer encoder that constructs a deep and contextualized representation of the full audio input with the aid of multi-head self-attention.

The most exciting idea of HuBERT is its training criterion, inspired by the masked language modeling task adopted by BERT on text. It works by concealing some of the input audio features and compelling the model to estimate the original, unmasked content. It does not predict the raw features, though, but rather the discrete "hidden unit" that the masked area belongs to.

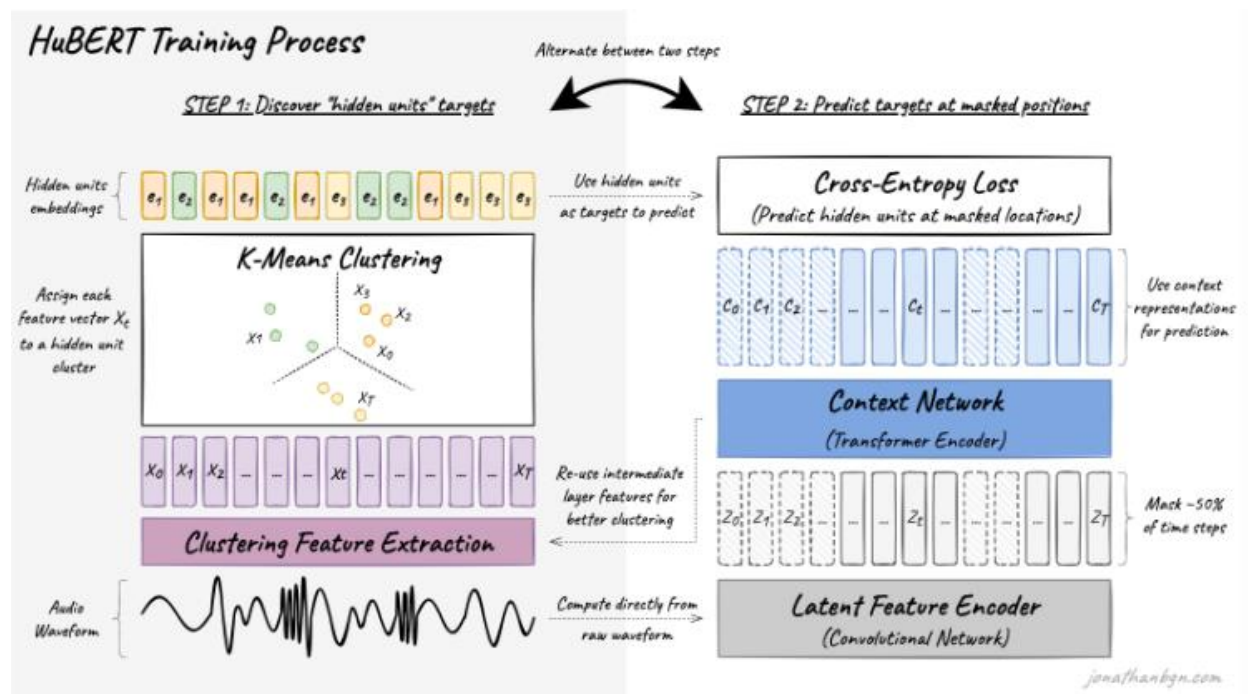


FIGURE 1: DISCRETE UNIT GENERATION (HUBERT + KMEANS)

HuBERT Clustering Step

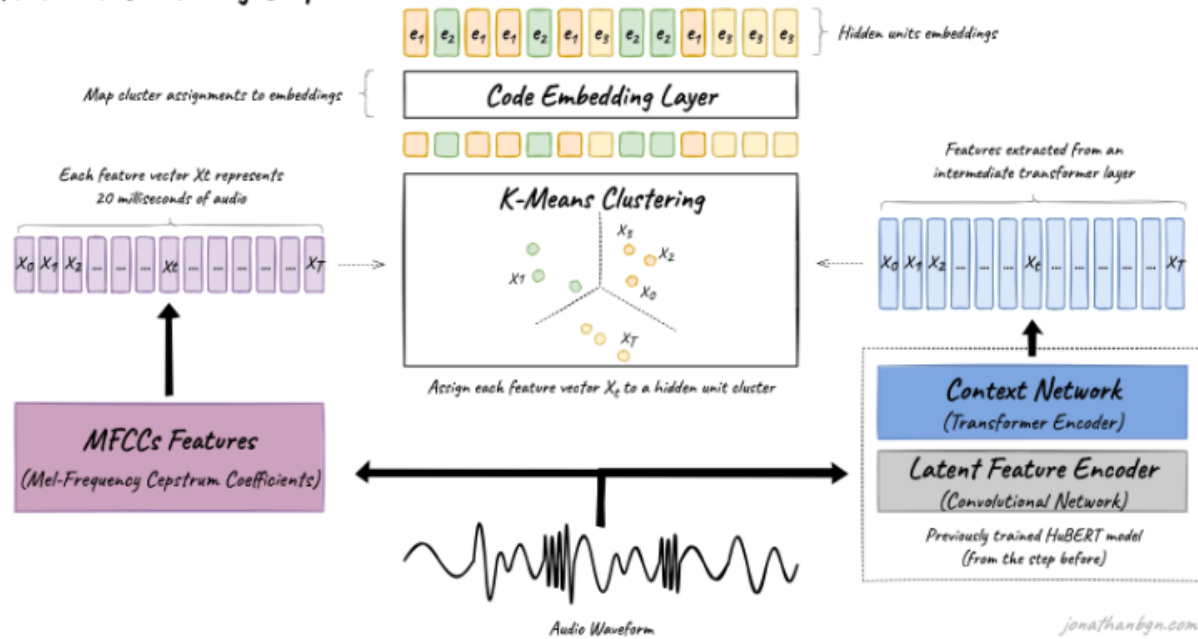


FIGURE 2: DISCRETE UNIT GENERATION 2 (HUBERT + KMEANS)

The model is optimized using a cross-entropy loss, yet importantly, the loss is computed only on the masked timesteps, represented by set M . This can be equated as:

as:

$$L_m(f; X, M, Z) = \sum_{t \in M} \log p_f(z_t | \bar{X}, t), \quad (1)$$

FIGURE 3: HUBERT CROSS ENTROPY EQUATION

Where:

- L_M is the loss over the masked steps.
- M is the set of time-step indices t that have been masked.
- X is the input speech sequence with the frames at indices M masked out.
- $p(z_t | \dots)$ is the model's predicted probability of the true discrete unit label z_t at the masked timestep t , given the surrounding unmasked context.

This has the effect of making the model learn a contextualized, high-dimensional meaning of the whole sequence of audio data in order to be able to predict the missing portions, rather than merely learn to reproduce its input. It is this that enables it to build strong, general-purpose representations of speech.

4.1.2 K-Means Clustering for Unit Discretization

According to the same paper [10], even the target discrete units (z_t) which the HuBERT model is trained to predict are, in turn, produced by a K-Means clustering procedure. This is done in two stages iteratively to produce the effective pre-trained model:

First Iteration: A basic K-Means model is initially trained on raw acoustic features (such as MFCCs) of the unlabeled data to produce an initial set of "teacher" labels. These labels are then used to train a first-pass HuBERT model to predict them.

Second Iteration (Refinement): The output of an intermediate layer of this trained HuBERT model is in turn fed as input to a new K-Means clustering procedure. This generates a significantly improved, stronger collection of discrete unit labels, which are afterwards utilized to instruct the final, strong HuBERT design.

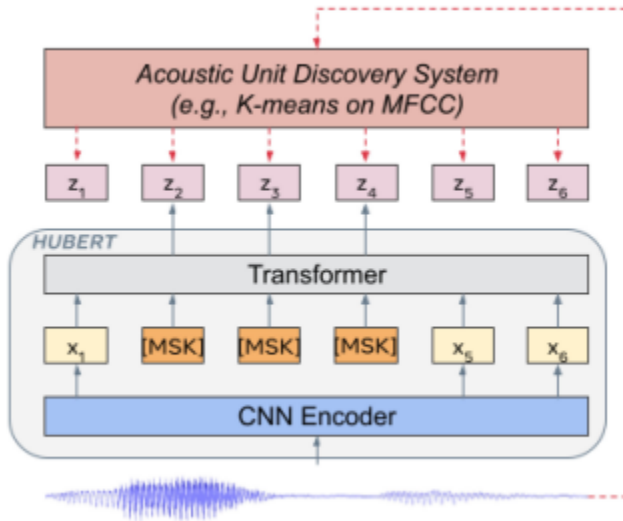


FIGURE 4: DISCRETE UNIT GENERATION 3 (HUBERT + KMEANS)

In the case of this thesis, a pre-trained HuBERT Base model, which already went through this process, was utilized. In order to construct the target vocabulary of our English target speech, the features of its 6th Transformer layer were extracted on our English training data and clustered with MiniBatchKMeans into the final 100-centers vocabulary.

4.2 Stage 2: Speech to Unit Translation Model

The main component of the translation pipeline is the S2UT model, that consists of the standard Transformer architecture. The particular model is the s2ut_transformer_fisher of the Fairseq toolkit. As it is represented in the study by Lee et al. [13], this type of architecture is aimed at directly projecting a source speech sequence into a series of discrete target units. It comprises of two primary parts which are an encoder and a decoder.

4.2.1 Encoder Architecture

The encoder is meant to take in the source speech spectrogram and construct a high-dimensional and contextualized representation of its content. At the start, it has a stack of 1D convolutional layers used as subsampling, where the input sequence length is reduced. This output is in turn passed into a stack of standard Transformer encoder layers, that incorporate multi-head self-attention to develop a deep context of the whole input utterance. More sophisticated speech encoders The standard Transformer layers are typically swapped with Conformer blocks [14] in state-of-the-art models such as Dub-S2ST to allow the network to attend to local acoustic details more effectively.

4.2.2 Decoder Architecture

The decoder is to use the output of the encoder and to autoregressively produce the target sequence of discrete English units. It consists of a stack of Transformer decoder layers, with each layer computing two kinds of attention: masked self-attention on the already generated units, and cross-attention on the output of the encoder to concentrate on the most interesting portions of the source audio. The activations of the last decoder layer are Linear projected and passed through a softmax to form a probability distribution over the 100-dimensional vocabulary, against which the next unit is sampled.

4.3 Stage 3: Unit-to-Speech Synthesis (Vocoder)

6.4.1 HiFi-GAN Vocoder Architecture The last step involves a pre-trained HiFi-GAN vocoder that decomposes the predicted sequence of discrete units into a high-quality, listenable speech waveform [15]. HiFi-GAN is a Generative Adversarial Network (GAN) model, which is very efficient and sounds very natural. Its Generator is a fully convolutional network with a Multi-Receptive Field Fusion (MRF) module to look at patterns on various scale simultaneously. The Discriminator of it is also based on a multi-scale structure to automatically detect and cancel artifacts in the synthesized audio effectively to achieve high fidelity. The pre-trained model selected in this project was because it is directly compatible with 100-cluster HuBERT units produced during our data preparation phase.

From: Benchmarking Hindi-to-English direct speech-to-speech translation with synthetic data

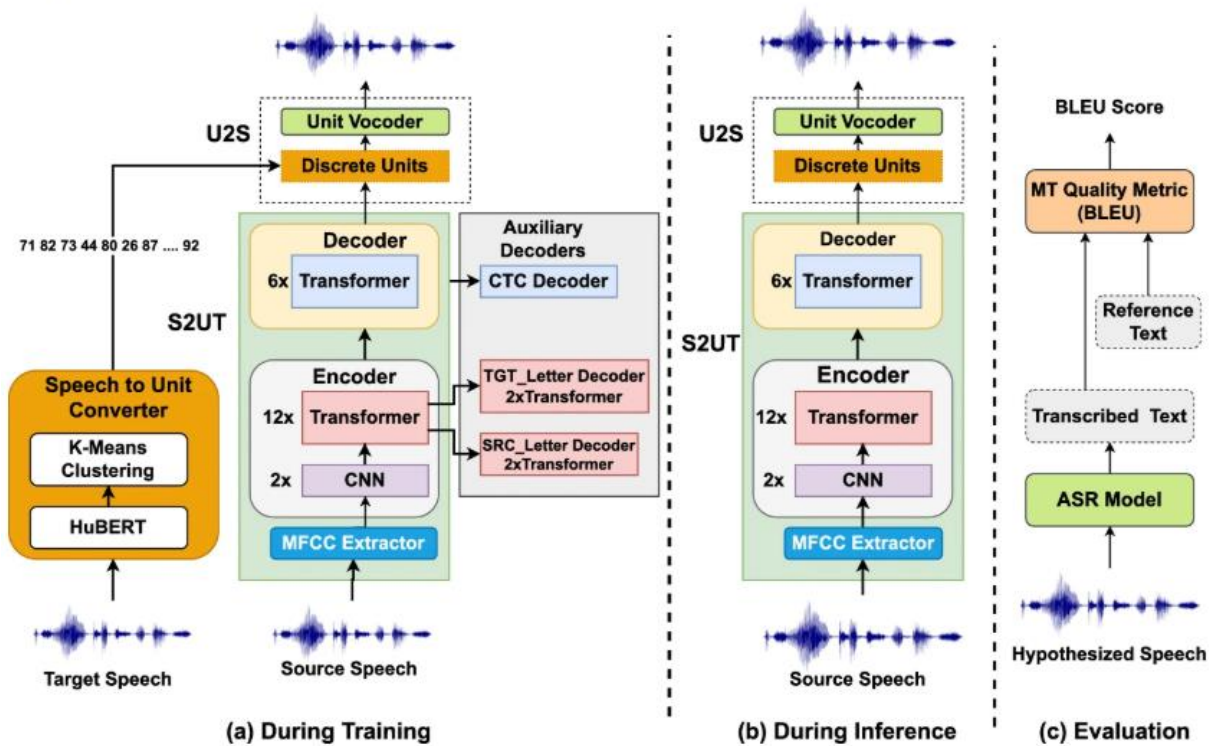


FIGURE 5: TEXTLESS SPEECH-TO-SPEECH ARCHITECTURE

5. Gantt Chart

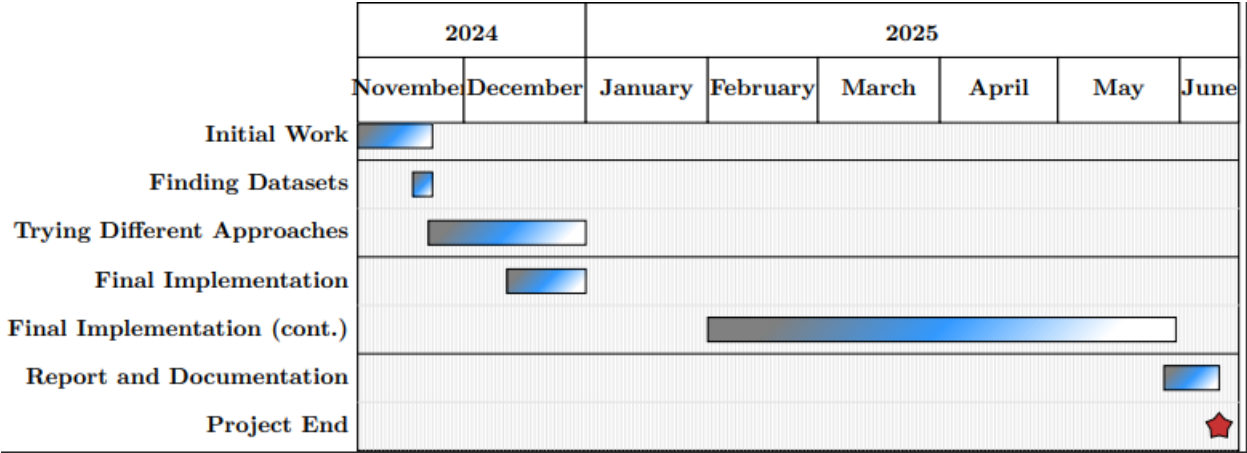


FIGURE 6: GANTT CHART

6. Implementation

The core of the project was to implement a fully textless speech-to-speech pipeline. That required significant data preparation, training, and customized evaluation workflow. The project was built in Python 3.10 with pytorch 2.1.0 on a cloud-based GPU.

6.1 Hardware and Platform Selection: From Colab to RunPod

Initially, development was on Google Colab Pro+, which has its own powerful GPU, NVIDIA A100 GPUs with 40GB of VRAM and 82GB of RAM. However, the platform showed a significant environment challenge as Colab's runtime uses Python 3.11 which is incompatible with Fairseq toolkit. Even after trying to change the Python version in Colab to 3.10, the library installation would fail, as its importation process appeared to maintain connections to the underlying system's default Python 3.11 libraries, creating irresolvable conflicts. It took 4 days to discover this.

To get around these environment limitations and obtain access to more powerful hardware for processing of a large CoVoST 2 dataset, the project has been moved to RunPod. was advantageous in many keyways.

Upgraded Hardware: They could use an NVIDIA A100 having 80GB VRAM and 125 GB RAM, which was very important for training large models with large batch sizes.

RunPod's persistent storage that persists like Local: which runs itself. It was a critical improvement because it got rid of having to mount Google Drive which was highly slowing down all data io operations I had during data preparation as well as during training.

Full control over the environment: choosing a dedicated instance to solve the problems of Python Version conflicts.

The tradeoff for this strategic shift to a more powerful platform and higher cost was to make sure that the project could succeed.

6.2 Environment Configuration & Reproducibility

The central framework of the main modeling tasks is Meta AI's Fairseq toolkit. It needed its robust implementation of the Transformer architecture to make the project. One of the major challenges in this project was to deal with library conflict and incompatibility. In order to resolve the long running issues with PryTorch, torchaudio, and the fairseq, the final working environment on the RunPod instance has been separated with the help of Conda 11.8 and Python 3.10.

6.3 Data Sourcing and Preparation

Much of the effort of the project was spent on sourcing, preparing and handling datasets suitable for any given experimental task. Choosing and aligning the data were central to the project as the project's findings would later show a strong dependency of data scale to the model's performance.

6.3.1 Dataset Selection and Sourcing

During the different phases of the work, we made use of three primary datasets

1. The first monolingual grammar correction task was based on the well-known LJSpeech dataset. This dataset comprises 60 hours of high-quality audio of a single female speaker reading passages from non-fiction books, containing 13,100 short audio clips. This was a clean baseline of a single speaker to develop the main parts of the S2UT pipeline.

2. In turn, the FLEURS (for Arabic-to-English) dataset was used for the first cross-lingual experiment. In order to run the training, first we aligned the source Arabic and the target English data by their common sample IDs and subsequently got a parallel corpus not too big, having only about 2,300 samples.

3. To test the hypothesis that data scale was the main limiter I assembled a much larger dataset for the German to English task, CoVoST 2 & CVSS. To this, a complex manual preparation occurred combining two different corpora.

The German audio used from the source was provided by Mozilla Common Voice Corpus 4.

CVSS-C corpus was used as the source of the target English audio and alignment files. A parallel speech–speech dataset of roughly 128,000 matching pairs of speech utterances in 5 languages was created by matching the samples of these two sources.

6.3.2 Storage and resource management

The size of the large scale CoVoST 2 / CVSS dataset was a feature resource challenge. The unpacked archives needed a large amount of persistent storage because their size combined. To train, storage was provisioned on the RunPod cloud platform at a recurring cost of \$2 USD per day that was an extra cost other than the cost of the GPU computing resources used for training. This brings to light a practical engineering tradeoff of large-scale model development: data management costs are as important as compute cost.

6.3.3 Workflow data processing

An important piece of the implementation was the pipeline to transform the raw, unpacked corpora into the correct format that Fairseq could use. Aligning the source and target data, and producing the discrete unit representation of the target speech were part of this. An example of this workflow happens to be the task of translating German to English using the CVSS and Common Voice datasets.

Once we get an official alignment file (train.tsv) which was released by the dataset CVSS-C, we start by reading the official alignment file using the pandas library. In this file, we have a mapping between source audio filename from Common Voice (i.e., a column named path) and target English translation text (i.e., another column named translation), which acts as ground truth.

Audio Processing & Intermediate Manifest Generation: The script then loops through each one of these entries in this alignment file. In each of the directories of source German audio and target English audio, there is source German audio and target English audio for every aligned pair. Both are loaded with torchaudio, resampled to 16kHz and then saved to its final destination for the source German audio. Within the process of processing the target English audio files, a temporary manifest file will be generated (Considering the fact that we keep the German source as WAV, and the English target is discrete). The path and duration of each resampled English speech audio file is provided in this manifest which is used as the direct input for the unit generation script.

Discrete Unit Generation: The manifest is then passed to the `quantize_with_kmeans.py` script for Discrete Unit Generation. The manifest files of English audio files are processed by this script using HuBERT, KMeans models and the result of a discrete unit sequence of an audio file is generated. It creates another intermediate file (e.g. `quantized_english_train.txt`), each line in which consists in an audio ID with suffix `_en`, then a pipe (`()`), then a space separated string of unit IDs.

```
Launcher x test.txt x manifest_english_test.txt +
1 /workspace/covost2_de_en_s2ut_data_final_prep/temp_english_wavs_for_s2u/test
2 common_voice_de_17299388_en.wav 39000
3 common_voice_de_17299389_en.wav 22000
4 common_voice_de_17299391_en.wav 28200
5 common_voice_de_17300032_en.wav 62000
6 common_voice_de_17300034_en.wav 46600
7 common_voice_de_17300036_en.wav 31800
8 common_voice_de_17300037_en.wav 38000
9 common_voice_de_17300137_en.wav 19200
10 common_voice_de_17300138_en.wav 31400
11 common_voice_de_17300139_en.wav 38800
12 common_voice_de_17300308_en.wav 10000
13 common_voice_de_17300310_en.wav 29000
14 common_voice_de_17300317_en.wav 49400
15 common_voice_de_17300426_en.wav 33200
16 common_voice_de_17300427_en.wav 21200
17 common_voice_de_17300431_en.wav 87800
18 common_voice_de_17300432_en.wav 32400
19 common_voice_de_17300433_en.wav 48000
20 common_voice_de_17300434_en.wav 40800
21 common_voice_de_17300435_en.wav 86200
22 common_voice_de_17300467_en.wav 73800
23 common_voice_de_17300468_en.wav 37600
24 common_voice_de_17300469_en.wav 36800
```

FIGURE 7: MANIFEST FILE FOR MATCHING THE DATA PAIRS (BEFORE DELETING THE SUFFIX _EN)

Final formatting the intermediate unit file: When each line is processed by a script, the sample ID is cleaned to fit the source audio ID (by stripping off the _en suffix), and the final and clean data is saved into the target unit files (train.txt, dev.txt, test.txt). The last step of input preparation would be to pass these files, consisting of the clean ID|unit_sequence pairs along with the directory of source German WAVs to S2UT.

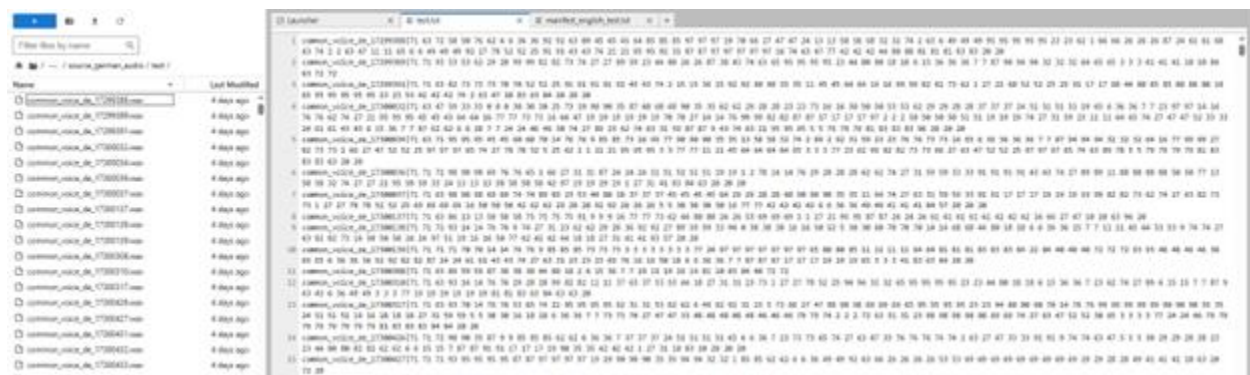


FIGURE 8: SOURCE GERMAN WAVs AND THE ENGLISH CORRESPONDING DISCRETE UNITS AFTER MATCHING

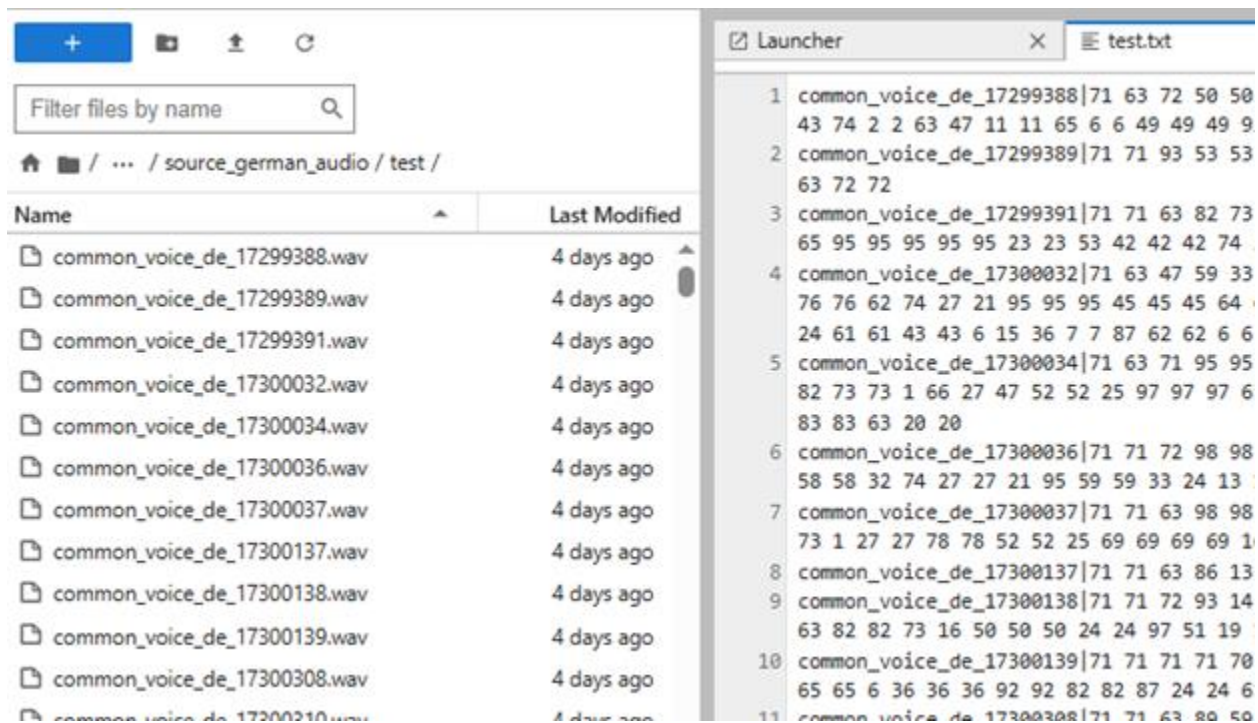


FIGURE 9: SOURCE GERMAN WAVS AND THE ENGLISH CORRESPONDING DISCRETE UNITS AFTER MATCHING (ZOOMED IN)

Filter files by name

Launcher

test.txt

test.tsv

manifest_english_test.txt

Delimiter: tab

		id	src_audio	src_n_frames	tgt_audio	tgt_n_frames
1	common_voice_de_17299388	_german_audio/test/common_voice_de_17299388.wav	319	91 43 74 21 95 92 31 87 97 16 74 63 47 77 42 44 80 81 83 20	73	
2	common_voice_de_17299389	_german_audio/test/common_voice_de_17299389.wav	223	3 65 95 23 44 80 18 6 15 36 7 87 94 32 64 65 3 41 10 84 63 72	43	
3	common_voice_de_17299391	_german_audio/test/common_voice_de_17299391.wav	285	8 44 80 85 88 80 18 65 95 23 53 42 74 2 63 47 10 83 63 84 20	54	
4	common_voice_de_17300032	_german_audio/test/common_voice_de_17300032.wav	633	27 89 23 62 74 63 31 59 87 9 43 74 63 21 95 5 79 81 83 96 20	116	
5	common_voice_de_17300034	_german_audio/test/common_voice_de_17300034.wav	482	9 82 73 66 27 63 47 52 25 97 65 74 63 89 78 5 79 81 83 63 20	94	
6	common_voice_de_17300036	_german_audio/test/common_voice_de_17300036.wav	331	2 74 27 21 95 59 33 24 13 58 42 97 19 1 27 31 41 83 84 63 20	59	
7	common_voice_de_17300037	_german_audio/test/common_voice_de_17300037.wav	427	69 16 50 42 62 29 28 92 26 5 30 16 77 42 6 36 49 41 84 57 20	65	
8	common_voice_de_17300137	_german_audio/test/common_voice_de_17300137.wav	333	4 80 26 53 69 1 27 21 95 87 24 61 42 16 66 27 47 10 83 96 20	33	
9	common_voice_de_17300138	_german_audio/test/common_voice_de_17300138.wav	326	82 73 16 50 24 97 51 19 16 50 77 42 44 18 27 31 41 83 57 20	66	
10	common_voice_de_17300139	_german_audio/test/common_voice_de_17300139.wav	489	7 63 31 23 69 76 16 50 18 6 36 7 87 17 19 65 3 41 83 63 84 20	59	
11	common_voice_de_17300308	_german_audio/test/common_voice_de_17300308.wav	309	71 63 89 59 87 38 44 80 18 2 6 15 36 7 19 81 10 83 84 40 72	21	
12	common_voice_de_17300310	_german_audio/test/common_voice_de_17300310.wav	614	62 74 27 89 6 15 7 87 9 43 6 36 49 3 77 19 81 83 63 84 63 20	60	
13	common_voice_de_17300317	_german_audio/test/common_voice_de_17300317.wav	494	3 31 23 98 69 74 27 63 47 52 30 65 3 77 24 46 79 81 83 84 20	86	
14	common_voice_de_17300426	_german_audio/test/common_voice_de_17300426.wav	470	3 44 80 82 62 6 15 7 87 91 17 19 90 35 42 1 27 31 10 83 10 20	64	
15	common_voice_de_17300427	_german_audio/test/common_voice_de_17300427.wav	278	1 85 62 6 36 49 92 63 66 26 53 69 29 28 49 41 10 63 20 72 20	31	
16	common_voice_de_17300431	_german_audio/test/common_voice_de_17300431.wav	938	11 65 3 77 98 69 50 11 45 64 1 85 5 79 29 6 28 49 41 83 10 20	157	
17	common_voice_de_17300432	_german_audio/test/common_voice_de_17300432.wav	472	23 1 66 63 47 24 58 65 6 36 7 87 9 16 77 42 80 81 83 63 83 20	59	

FIGURE 10: FINAL .TSV FILE THAT INCLUDES SOURCE GERMAN SPEECH WITH CORRESPONDING ENGLISH DISCRETE UNITS BEFORE BEING PASSED TO S2UT

Audio Processing and Speech Synthesis Workflow

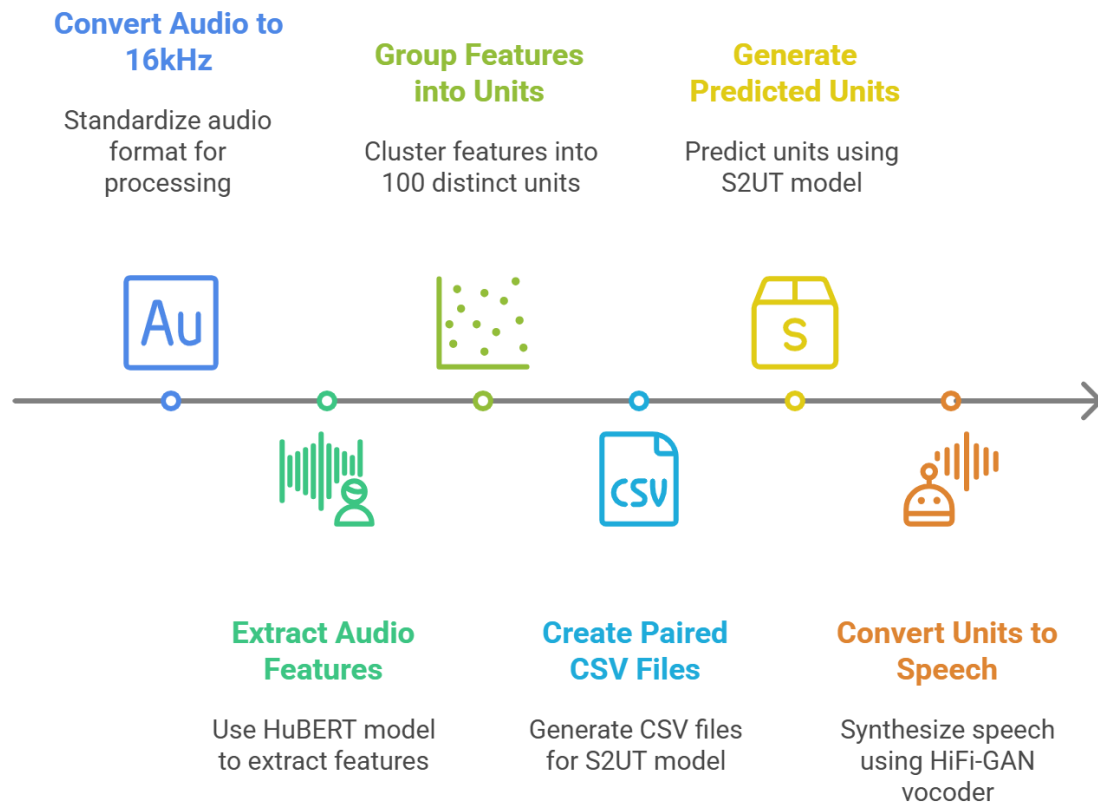


FIGURE 11: SUMMARIZED SPEECH WORKFLOW

6.4 ASR Based Transcription

The synthesized speech output of a **HiFi-GAN vocoder** from the discrete units predicted by the S2UT model, was transcribed back to text through an ASR model which was powerful and pre-trained. For such a task we selected the Whisper model ([openai/whisper-medium.en](https://openai.github.io/whisper-medium.en)) because of its high zero-shot accuracy. This hypothesis text is then compared against a ground-truth reference text using a variety of standard NLP metrics. Nevertheless, its output displayed some aberrations as far as numbers were concerned, for instance if you said 'three' it would transcribe the same as the digit '3', or in other cases it misunderstood number words, like converting 'eighty' to 'at'. Other powerful models such as NVIDIA's ParakeetTDT were investigated; but were not included as they are huge and require very high computational resources.

7. Evaluation Methodology & Metrics

Evaluating the performance of textless speech-to-speech translation model brings forth a peculiar challenge: as the model generates speech waveforms but never actual text, a comparison cannot be drawn against a text reference. This chapter sheds light on the multi-stage pipeline and metric suite by which quantitative and qualitative assessments of the built models have been made during this project.

7.1 Training-Based Metrics (loss vs. nll_loss)

Discussed below are the two losses during training, which gave us the model's rates of convergence.

NLL Loss (Negative Log Likelihood Loss): This is the pure cross-entropy loss value, used as a direct measure of the capacity of the model to predict. For one correct target unit c , is calculated as

$$L_{NLL} = -\log(p_c) \quad (2)$$

where P_c is the probability the model assigned to the correct unit. A low `nll_loss` indicates that the model, in general, is more confident in its prediction-and in this sense, it is a better prediction.

Loss: This is the value the optimizer uses, adding in the label smoothing penalty as proposed by Szegedy et al. (2016). Label smoothing is a form of regularization that discourages the model from giving full probability to the correct class. The full loss is given by

$$L_{smooth} = (1 - \epsilon) * L_{NLL} + \epsilon * \left(\frac{1}{K} \right) * \sum_{i=1}^K -\log(p_i) \quad (3)$$

where ϵ is a smoothing factor (e.g., 0.1 or 0.2), and K is the total number of units. Thus, full loss is always greater than `nll_loss`.

7.2 Text-Based Evaluation Metrics

Once the output speech was transcribed into text, the evaluation was carried out based on the following metrics:

BLEU (Bilingual Evaluation Understudy): It is the most common metric in machine translation, but it essentially measures the precision of the n -grams (word sequences) in the hypothesis compared to the reference, punishing translations that are too short.

BLEU measures n -gram precision with a brevity penalty: $BLEU = BP \times \exp\left(\sum_{n=1}^N w_n \log p_n\right)$ (3)

where:

P_n is the precision of n -grams (typically for $n = 1$ to 4).

w_n are weights for each n -gram precision (usually uniform, e.g., 0.25 for each).

BP is the "Brevity Penalty," which penalizes generated translations that are shorter than the reference. It is calculated as:

$$BP = \begin{cases} 1, & c \leq r \\ e^{\{1-r-c\}}, & c > r \end{cases}$$

where c is the length of the hypothesis and r is the length of the reference

chrF (character F-score): This metric operates at the character level, comparing character n-grams (sequences of characters) between the hypothesis and the reference. By doing so, it is naturally robust to out-of-vocabulary words, misspellings, and morphological variations (e.g., 'running' vs 'run'), which can unfairly penalize word-level metrics like BLEU. It calculates an F-score based on character n-gram precision and recall. A higher score is better.

$$chrF = 2 * \frac{(precision * recall)}{(precision + recall)} \quad (4)$$

where precision and recall are based on character n-gram overlaps.

WER (Word Error Rate): This is the standard measure for ASR. The number of substitutions, deletions, and insertions required to transform the hypothesis text into the reference text is calculated. The lower the score, the better.

$$WER = \frac{(S + D + I)}{N} \quad (5)$$

where S = substitutions, D = deletions, I = insertions, and N = reference word count.

TER (Translation Edit Rate): TER is considered to be more suitable for translation in comparison to WER, as it treats shifting a block of words, which is correct in itself, to a different position as a separate "edit" apart from substitutions, insertions, and deletions. The lower score is better.

$$TER = \frac{(Edits)}{(\text{Reference Length})} \quad (6)$$

where edits include substitutions, insertions, deletions, and shifts.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation): ROUGE is typically used in text summarization and measures n-gram recall, with the eye on how many words in the reference appear in the hypothesis, thus making it not so strict about word order as in BLEU. Hence, the higher the score, the better.

$$ROUGE - N = \frac{(\sum_{\{n \text{ in Reference } N\text{-grams}\}} \text{Count}_{\text{match}(n)})}{(\sum_{\{n \text{ in Reference } N\text{-grams}\}} \text{Count}(n))} \quad (7)$$

METEOR (Metric for Evaluation of Translation with Explicit Ordering): An advanced metric that performs alignment between hypothesis and reference not only based on exact word matches but also on their stems and synonyms (via WordNet). This renders it more semantically oriented than BLEU or WER. A higher score is better.

$$METEOR = F_{\{mean\}} \cdot (1 - P_{\{penalty\}}) \quad (8)$$

where $F_{mean} = 10 * \frac{(P * R)}{(P + 9 * R)}$ P = precision, R = recall, and $P_{penalty}$ accounts for word order

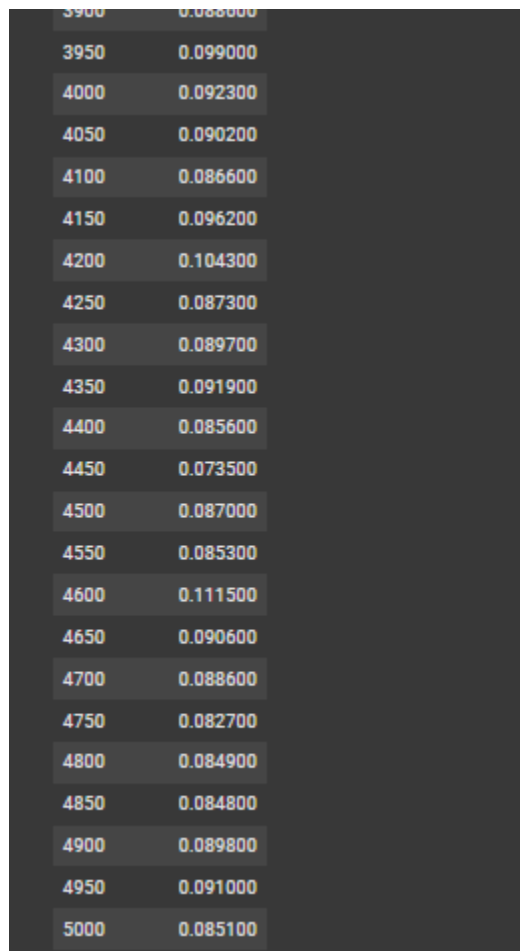
8. Trials, Results, and Discussion

The core experimental work in the project is presented in this chapter. The experiments were designed as a sequence, whereby the findings of one informed the methodology of the next. To be able to translate the speech starting from nothing, we started from foundational experiments to overcome architectural

challenges, a baseline monolingual task to validate the pipeline, and finished by successfully deploying a large, cross-lingual speech to speech translation model.

8.1 Initial Trials: Overcoming Architectural Incompatibilities

The first hypothesis with the project was that a more expressive acoustic vocabulary (one that included more phonetic detail) would produce a higher fidelity result. An initial experiment was then run with a HuBERT model to get a 1000 cluster unit vocabulary. We then successfully fine-tuned a text-based BART model to perform grammar correction on these 1000-unit sequences directly and got promising results in the unit-to-unit level.



3900	0.088000
3950	0.099000
4000	0.092300
4050	0.090200
4100	0.086600
4150	0.096200
4200	0.104300
4250	0.087300
4300	0.089700
4350	0.091900
4400	0.085600
4450	0.073500
4500	0.087000
4550	0.085300
4600	0.111500
4650	0.090600
4700	0.088600
4750	0.082700
4800	0.084900
4850	0.084800
4900	0.089800
4950	0.091000
5000	0.085100

FIGURE 12: INITIAL TRIAL - BART LOSS

But finally, this approach was blocked at the stage of final synthesis. The key challenge was that a compatible HiFi-GAN vocoder needed to be trained with the same number (1000) of clusters and using the same HuBERT layer features as used for unit generation. There existed other 1000 cluster vocoders but for different HuBERT layers hence they were incompatible.

Despite the originally intended approach, an alternative workaround was explored in an attempt to salvage the process. Instead of a direct vocoder, after training an acoustic model translating Unit to Mel, they used the output during training as labels to generate device specific Unit Hierarchies and Unit Recyclers to generate WaveNet input. This is the mel-spectrogram from the BART model and this UnitToMelTransformer modifies this "corrected" unit sequence to create a mel-spectrogram. However,

such spectrogram can further be processed through standard HiFi-GAN vocoder to synthesize the final waveform. After several days of training, this model made audio that was significantly better than a random baseline, though still noisy and lacking in the necessary clarity.

The early results could have been touted as a partial success; however, because the focus of the project was to finish a high-fidelity, user-ready product, this work was discarded in favor of a stronger pipeline utilizing a 100-unit vocabulary for which compatible and high-quality vocoders could be found.

8.2 Trial 1: Monolingual Grammar Correction

Making the component pipeline stable, the first major experiment concerned monolingual English-to-English grammar correction. This was a very useful baseline for establishing whether the S2UT architecture really is capable of learning complicated sequence-to-sequence transformations on speech.

8.2.1 Approach: A Novel Solution for Non-Paired Data

Due to the lack of any grammar speech pairs, the synthetic corpus was created by corruption of the LJSpeech dataset. After conversion of clean audio into a "correct" discrete unit representation, the corruption function introduced various errors into corresponding discrete unit streams like unit duplication, deletion, and reordering. To be very specific, in harmony with the essence of the S2UT model, the corrupted discrete units were then vocoded by HiFi-GAN to produce audible "incorrect" speech. This synthesized speech thereby served as the source input, forming a large training corpus with perfect alignment: (incorrect speech WAV, correct unit sequence).

8.2.2 Evaluation: A Multistage Strategy

An evaluation of the model required a careful and multi-stage approach to take care of the complications in the ASR-based pipeline alongside the nature of the LJSpeech dataset.

Stage 1: Baseline Evaluation on Raw Text

Initially, an evaluation was conducted treating the ASR output as raw text against the raw reference text. The traditional metrics, such as BLEU and WER, gave scores (67.86 and 24, respectively) because they penalized superficial aspects such as casing and punctuation, which do not the model's core grammatical correction capabilities.

```
[
{
  "name": "BLEU",
  "score": 67.8622,
  "signature": "nrefs:1|case:lc|eff:no|tok:13a|smooth:exp|version:2.5.1",
  "verbose_score": "84.4/72.8/64.1/56.6 (BP = 0.988 ratio = 0.988 hyp_len = 24864 ref_len = 25162)",
  "nrefs": "1",
  "case": "lc",
  "eff": "no",
  "tok": "13a",
  "smooth": "exp",
  "version": "2.5.1"
},
{
  "name": "chrF2",
  "score": 82.3817,
  "signature": "nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.5.1",
  "nrefs": "1",
  "case": "mixed",
  "eff": "yes",
  "nc": "6",
  "nw": "0",
  "space": "no",
  "version": "2.5.1"
},
{
  "name": "TER",
  "score": 21.8838,
  "signature": "nrefs:1|case:lc|tok:tercom|norm:no|punct:yes|asian:no|version:2.5.1",
  "nrefs": "1",
  "case": "lc",
  "tok": "tercom",
  "norm": "no",
  "punct": "yes",
  "asian": "no",
  "version": "2.5.1"
}
]

Calculating ROUGE scores...
ROUGE1 F1 Score: 0.8515
ROUGE2 F1 Score: 0.7633
ROUGE1 F1 Score: 0.8509
METEOR Score: 0.7847
```

FIGURE 13: TRAIL 1 , GRAMMAR CORRECTION BLEU, CHRf2, TER, ROUGE, AND METEOR SCORES ON RAW TEXT (BEFORE NORMALIZATION)

Showing 5 inference examples:

Example 1:
Reference: but Oswald had rented post office box three zero zero six one in New Orleans on June three, nineteen sixty-three,
Hypothesis: but Oswald had rented Post Office Box 304061 in New Orleans on June 3, 1963.

Example 2:
Reference: However, there is no evidence that these men failed to take any action in Dallas within their power that would have averted the tragedy.
Hypothesis: However, there is no evidence that these men failed to take any action in Dallas with their power that would have averted the tragedy. As will be seen,

Example 3:
Reference: At one time the Marshalsea was the receptacle of pirates, but none were committed to it after seventeen eighty-nine.
Hypothesis: At one time, the Martian sea was the wristicle of pirates. But none were committed to it. Act 70. At Elie.

Example 4:
Reference: Remainder of motorcade.
Hypothesis: in the nature of motorcade.

Example 5:
Reference: and with all the appearances of spontaneity as locomotive bodies,
Hypothesis: and with all the appearances of spontaneity as locomotive buddies.

FIGURE 14: TRAIL 1 GRAMMAR CORRECTION - INFERENCE ON RAW TEXT (BEFORE NORMALIZATION)

Example 6:
Reference: shaking her clenched and manacled hands in the officers' faces.
Hypothesis: shaking her clinched and manacled hands in the officer's faces.

Example 7:
Reference: In the period from November nineteen sixty-one to November nineteen sixty-three,
Hypothesis: in the period from November 1961 to November 1963.

Example 8:
Reference: and there seems to have been good reason for supposing that he was a greater villain than any of those arraigned.
Hypothesis: and there seems to have been good reason for supposing that he was a greater villain than any of those arraigned.

FIGURE 15: TRAIL 1 GRAMMAR CORRECTION - INFERENCE 2 ON RAW TEXT (BEFORE NORMALIZATION)

Stage 2: Full Normalization and Its Limitations

To get a fairer comparison, a "full normalization" pipeline was developed. This process converted all text to lowercase, removed all punctuation, and attempted to convert number words to digits (e.g., "eight" to "8"). This full normalization significantly improved scores across the board, with the BLEU score, for example, increasing to **74.27**. However, a deeper analysis revealed that this number conversion step was flawed. As shown in the side-by-side diff analysis (see **Figure 17-20**), the normalization script struggled with the complex dates and multi-word numbers found in the LJSpeech data. This meant that while the scores were higher, they were based on a comparison where the reference text itself had been partially corrupted by the imperfect normalization process.

```
[
{
  "name": "BLEU",
  "score": 74.2691,
  "signature": "nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.5.1",
  "verbose_score": "86.6/78.3/71.4/65.4 (BP = 0.990 ratio = 0.990 hyp_len = 22129 ref_len = 22345)",
  "nrefs": "1",
  "case": "mixed",
  "eff": "no",
  "tok": "13a",
  "smooth": "exp",
  "version": "2.5.1"
},
{
  "name": "chrF2",
  "score": 87.769,
  "signature": "nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.5.1",
  "nrefs": "1",
  "case": "mixed",
  "eff": "yes",
  "nc": "6",
  "nw": "0",
  "space": "no",
  "version": "2.5.1"
},
{
  "name": "TER",
  "score": 16.6883,
  "signature": "nrefs:1|case:lc|tok:tercom|norm:no|punct:yes|asian:no|version:2.5.1",
  "nrefs": "1",
  "case": "lc",
  "tok": "tercom",
  "norm": "no",
  "punct": "yes",
  "asian": "no",
  "version": "2.5.1"
}
]

Calculating ROUGE and METEOR scores...
ROUGE1 F1 Score: 0.8545
ROUGE2 F1 Score: 0.7667
ROUGEL F1 Score: 0.8539
METEOR Score: 0.8438
```

FIGURE 16: TRAIL 1 , GRAMMAR CORRECTION BLEU, CHRF2, TER, ROUGE, AND METEOR SCORES ON FULLY NORMALIZED TEXT

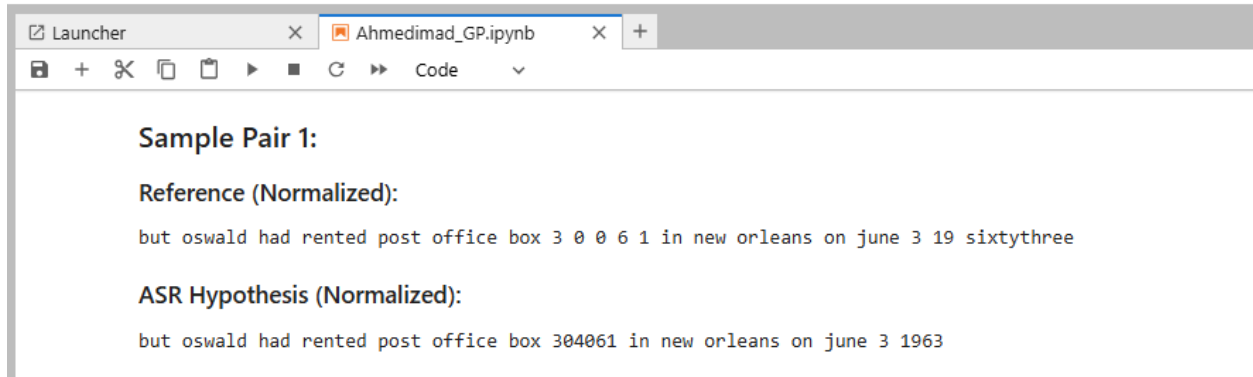


FIGURE 17: TRAIL 1 GRAMMAR CORRECTION - INFERENCE 2 ON FULLY NORMALIZED TEXT

Side-by-Side Word-Level Diff:

	Reference Words			Hypothesis Words	
f	1	but	f	1	but
	2	oswald		2	oswald
	3	had		3	had
	4	rented		4	rented
	5	post		5	post
	6	office		6	office
	7	box		7	box
n	8	3	n	8	304061
	9	0			
	10	0			
	11	6			
	12	1			
	13	in		9	in
	14	new		10	new
	15	orleans		11	orleans
	16	on		12	on
	17	june		13	june
	18	3		14	3
t	19	19	t	15	1963
	20	sixtythree			

FIGURE 18: TRAIL 1 GRAMMAR CORRECTION - INDEX MATCHING CHECK ON FULLY NORMALIZED TEXT

Sample Pair 7:

Reference (Normalized):

in the period from november 19 sixtyone to november 19 sixtythree

ASR Hypothesis (Normalized):

in the period from november 1961 to november 1963

Side-by-Side Word-Level Diff:

	Reference Words			Hypothesis Words	
f	1	in	f	1	in
	2	the		2	the
	3	period		3	period
	4	from		4	from
	5	november		5	november
n	6	19	n	6	1961
	7	sixtyone			
	8	to	7	7	to
	9	november		8	november
t	10	19	t	9	1963
	11	sixtythree			

FIGURE 19: TRAIL 1 GRAMMAR CORRECTION - INDEX MATCHING CHECK 2 ON FULLY NORMALIZED TEXT

Sample Pair 5:

Reference (Normalized):

and with all the appearances of spontaneity as locomotive bodies

ASR Hypothesis (Normalized):

and with all the appearances of spontaneity as locomotive buddies

Side-by-Side Word-Level Diff:

	Reference Words			Hypothesis Words	
f	1	and	f	1	and
	2	with		2	with
	3	all		3	all
	4	the		4	the
	5	appearances		5	appearances
	6	of		6	of
	7	spontaneity		7	spontaneity
	8	as		8	as
	9	locomotive		9	locomotive
t	10	bodies	t	10	buddies

FIGURE 20: TRAIL 1 GRAMMAR CORRECTION - INDEX MATCHING CHECK 3 ON FULLY NORMALIZED TEXT

Stage 3: Partial Normalization and a More Robust Metric Suite

To resolve this, a final evaluation was conducted using a **partial normalization** approach. This involved lowercasing the text and removing punctuation but **critically excluded the imperfect number-to-digit conversion step**. While this yielded slightly lower scores than the full normalization, it is considered the most reliable and crucial measure of performance as it does not corrupt the contextual meaning between the reference and the hypothesis.

Example 28:

Reference: nothing to equal the excitement caused by the forgeries of robert ferdinand pries had been known before in the city of london

Hypothesis: nothing to equal the excitement caused by the forgeries of robert ferdinand preece had been known before in the city of london

Example 29:

Reference: it was so used for the date of the abolition of the star chamber in the sixteenth charles the first

Hypothesis: it was so used for the date of the abolition of the star chamber in the 16th charles i

Example 31:

Reference: 1 cannot estimate when prior to november 20 oswald made the paper bag

Hypothesis: 1 cannot estimate when prior to november 22 oswald made the paper bag

Example 32:

Reference: he would shoot any man or any policeman like a dog or any number of them who had treated him in that way

Hypothesis: he would shoot any man or any policeman like a dog or any number of them who had treated him in that way

FIGURE 21: TRAIL 1 GRAMMAR CORRECTION - INFERENCE ON PARTIALLY NORMALIZED TEXT (THE SUCCESS STORY)

```

L
{
  "name": "BLEU",
  "score": 73.8268,
  "signature": "nrefs:1|case:mixed|eff:no|tok:13a|smooth:exp|version:2.5.1",
  "verbose_score": "86.3/77.9/70.9/64.8 (BP = 0.990 ratio = 0.990 hyp_len = 22129 ref_len = 22345)",
  "nrefs": "1",
  "case": "mixed",
  "eff": "no",
  "tok": "13a",
  "smooth": "exp",
  "version": "2.5.1"
},
{
  "name": "chrF2",
  "score": 86.3125,
  "signature": "nrefs:1|case:mixed|eff:yes|nc:6|nw:0|space:no|version:2.5.1",
  "nrefs": "1",
  "case": "mixed",
  "eff": "yes",
  "nc": "6",
  "nw": "0",
  "space": "no",
  "version": "2.5.1"
},
{
  "name": "TER",
  "score": 16.9479,
  "signature": "nrefs:1|case:lc|tok:tercom|norm:no|punct:yes|asian:no|version:2.5.1",
  "nrefs": "1",
  "case": "lc",
  "tok": "tercom",
  "norm": "no",
  "punct": "yes",
  "asian": "no",
  "version": "2.5.1"
}
]

Calculating ROUGE and METEOR scores...
ROUGE1 F1 Score: 0.8516
ROUGE2 F1 Score: 0.7623
ROUGEL F1 Score: 0.8510
METEOR Score: 0.8436

```

FIGURE 22: TRAIL 1 GRAMMAR CORRECTION - BLEU, CHRf2, TER, ROUGE, METEOR SCORES ON PARTIALLY NORMALIZED TEXT (THE SUCCESS STORY)

WER: 0.1799

FIGURE 23: TRAIL 1 GRAMMAR CORRECTION - WER SCORE ON PARTIALLY NORMALIZED TEXT (THE SUCCESS STORY)

8.2.3 Adopting the Final Evaluation Metrics

As BLEU, WER and TER were not suitable because of their rigorous lexical matching, the metrics used in the final evaluation were those that may be more appropriate to measure the performance depending on the nature of the task.

8.2.3.1 A Critical View of METEOR and chrF2

First, METEOR (Metric for Evaluation of Translation with Explicit Ordering) seemed to be a perfect solution. Its advantages lie in the fact that it goes beyond the simple word matching, as it uses a WordNet database to match synonyms as well as the fact that an ASR output of the number 1 is semantically identical to a reference of the number one. This solves the number-word problem that afflicted BLEU.

Nevertheless, one crucial weakness was found regarding this particular grammar correction exercise: the use of stemming in METEOR as well as in chrF2. They see words of the same root, like running and run, as a match. This has a direct impact on the intent of the grammar correction model which is none other than to correct wrong verb tenses and forms of words. Failure to penalize such errors would give an inflated and false score by METEOR. Thus, although the METEOR score was high (~84.4), it could not be assumed that it was a valid indicator of the grammatical correction ability of the model.

8.2.4.2 ROUGE as the Most Confident Indicators

Having shown that both BLEU, WER, TER, chrF2 and METEOR were unsuitable, the final consideration was based on ROUGE as the most suitable matrix in this special task:

ROUGE: The reason this metric was selected is that recall of n-grams (content overlap) is less harsh to small word reordering and grammatical function words, which gives a better indication of the overall content recovery. It has offered a good ROUGE-L (85.10) score.

8.2.4.3 Trail 1's Final Results and Discussion

(Scale from 1 to 100)

Matrix	Raw / Validity	Full Normalization / Validity	Partial Normalization / Validity
BLEU	~67.86 / Invalid	~74.27 / Invalid	~73.83 / Valid but could've been higher
chrF2	~82.38 / Invalid	~87.8 / Invalid	~86.31 / Invalid
TER	~21.88 / Invalid	~16.7 / Invalid	~16.9 / Invalid
WER	~24 / Invalid	NAN / Invalid	17.99/ Valid but could've been better
METEOR	~78.85 / Invalid	~84.4 / Invalid	~84.4 / Invalid
ROUGEL	~85.09 / Invalid	~85.39 / Invalid	~85.10 / Valid

FIGURE 24: TRAIL 1 GRAMMAR CORRECTION - FINAL RESULTS COMPARISON TABLE

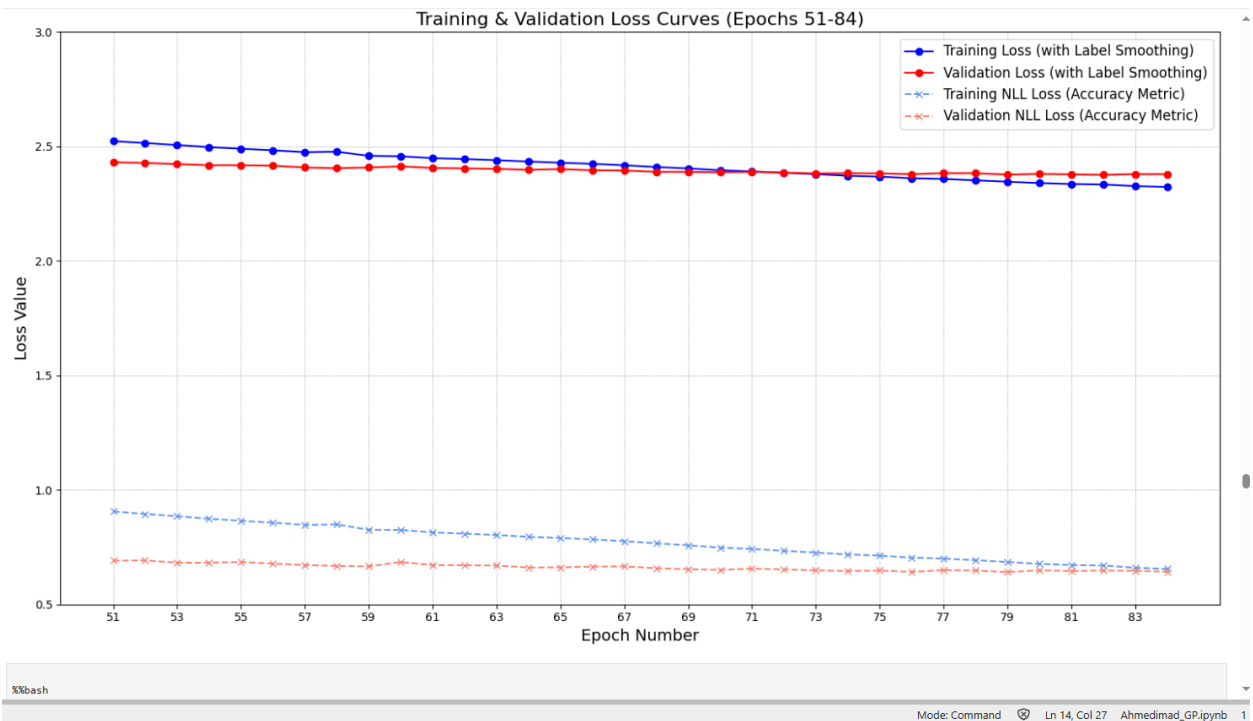


FIGURE 25: TRAIL 1 GRAMMAR CORRECTION - LOSS AND NLL_LOSS CURVE

8.2.4.3 Discussion

As the application of a textless S2UT pipeline to monolingual grammar correction is a novel approach, there was no existing speech-based model for direct comparison. Therefore, the objective of this evaluation was to establish a strong first baseline.

The final results are highly significant. The high BLEU score of **~73.83** and low WER of **~17.99%** after normalization demonstrate the viability of this approach. They prove that a Transformer-based S2UT model can successfully learn the complex patterns of grammatical structures from audio alone and perform corrections in a textless manner. This work provides a strong benchmark that future research in this new domain can build upon.

8.3 Trial 2: Low-Resource Speech Translation (Arabic-to-English)

8.3.1 Setup and Rationale

The second experiment was set as a direct experiment to test the S2UT pipeline on a cross-lingual task and to investigate its dependence on the size of the training data. This experiment selected the FLEURS dataset, which has parallel data in Arabic-to-English translation. The source and target speech were aligned based on their shared IDs, after which the obtained training corpus appeared to be quite small, comprising merely ~2,300 samples. This was enough to test the functionality of the pipeline, but it was speculated that it would not be enough to make the model learn the high-dimensional mapping needed to achieve high-quality translation.

8.3.2 Results and Discussion: A Successful Proof-of-Concept

As suspected, the model did not learn well using the little data. The loss decreased until the training process converged soon but with a relatively a bit high validation loss and nll_loss. This was corroborated by a qualitative examination of the synthesized English speech output; the audio was noticeably noisy,

and the produced sentences tends to be repetitive and semantically unrelated to the original Arabic speech.

Though it did not result in a practical model of translation, this was an important and successful proof of concept as even the loss is high, but it significantly decreased, starting from 9 till reaching almost 3.5, as shown in the loss curve below. It gave compelling evidence that the S2UT architecture, though adequate, is very demanding on large-scale datasets to reduce the complexity of the cross-lingual speech-to-speech translation. The discovery directly prompted the following attempt.

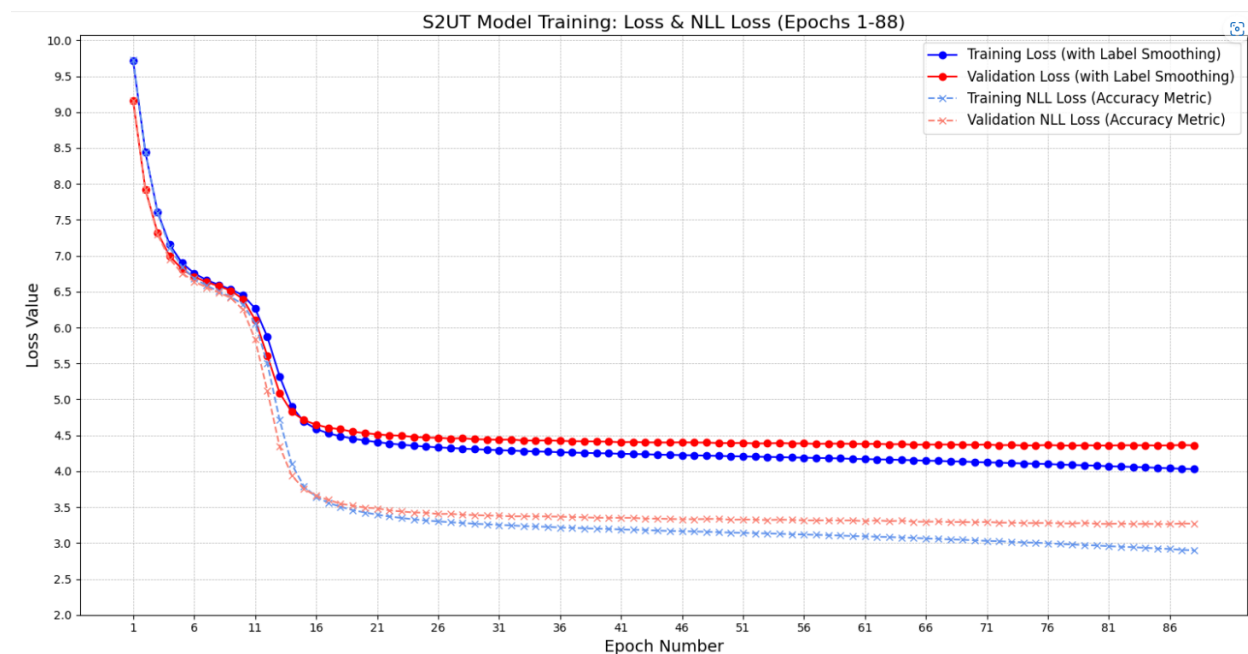


FIGURE 26: TRIAL 2 ARABIC TO ENGLISH TRANSLATION - LOSS AND NLL_LOSS CURVE

8.4 Trial 3: High-Resource Speech Translation (German-to-English)

The goal of this last experiment was to confirm the hypothesis of the previous one by running the same S2UT pipeline on an extremely large dataset.

8.4.1 A Macro Level Approach

In the case of this experiment a large-scale German-to-English parallel speech corpus was compiled, consisting of two parts: the source German audio was obtained via the Mozilla Common Voice Corpus 4 and the corresponding target English synthesized audio was selected via the CVSS-C corpus. This generated a quality-aligned training dataset of about 128000 samples. They then used the same S2UT pipeline: Preprocessing the source German audio, and converting the target English audio into a 100-cluster discrete unit vocabulary Hubert+KMeans pipeline. The very size of such data became a serious problem of resources, and a cloud-based GPU with a large storage was needed. Compared to earlier trails, the model consumed a lot of time during training because the size of the datasets was over 190GB.

8.4.2 Training Performance and Qualitative Analysis

Training on such a large dataset radically changed the performance of the model compared to the low-resource experiment. Training was stable and the model converged to nll_loss lower than the past Arabic-English experiment by almost 1.3.

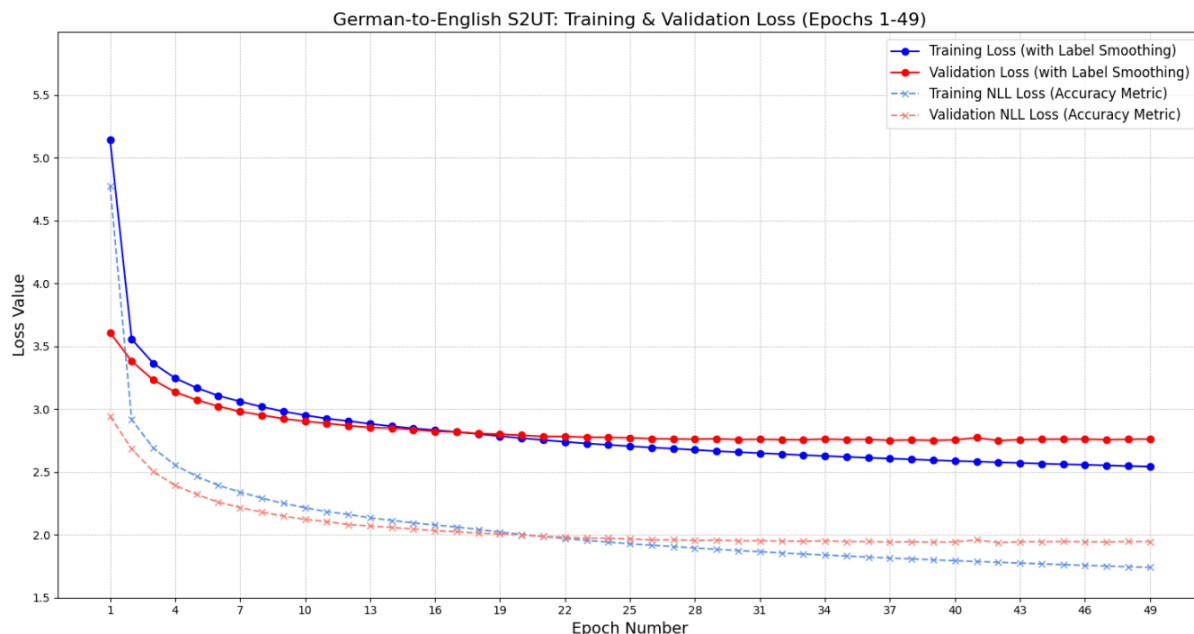


FIGURE 27: TRAIL 3 GERMAN TO ENGLISH TRANSLATION - LOSS AND NLL_LOSS CURVE

A qualitative analysis of the synthesized output revealed a significant breakthrough. Unlike the low-resource Arabic-to-English trial where the model failed to learn meaningful patterns, resulting in noisy output, increasing the data to 128,000 samples allowed the model to learn the patterns of the target language effectively. The resulting output voice was good and very clear.

However, while the acoustic fluency was high, the model sometimes struggles and gives wrong translation. This highlights the immense difficulty of the end-to-end S2ST task. This is a very complex task, especially when it is about two different languages. The model needs to not only master the acoustic characteristics of both languages but also bridge the vast semantic gap between them. It must understand the source German speech and analyze and generate the corresponding target discrete units in the another language.

The presence of these translation errors, even with a big dataset of 128k records, suggests that achieving near-perfect semantic accuracy with this architecture likely needs a larger dataset consisting of millions of records. This points to a significant bottleneck in the field, as we suffer from the lack of such massive, publicly available parallel speech corpora. Furthermore, the move to a larger dataset introduced the additional challenge of computational power, which added complexity to the project after the Arabic experiment. It is important to note that this is a common challenge in the S2ST literature, and many state-of-the-art other papers suffer from this and give lower BLEU scores compared to text-based translation, reflecting the inherent difficulty of the task .

8.4.3 Evaluation: A Comprehensive and Reliable Assessment

Given the model's performance, a full suite of metrics was used for evaluation. The text from the CoVoST 2 dataset is transcribed speech and does not contain the same high density of complex numbers and dates

that made the LJSpeech data problematic. Therefore, all evaluation metrics are considered reliable for this task.

Before final scoring, a text normalization process was applied to both the ASR hypothesis and the reference text. This involved handling the upper case and removing commas, dots, question marks, etc. As expected, this normalization significantly changed the BLEU score positively, allowing for a fairer comparison of the core word content.

Before normalization

```
--- Calculating BLEU, CHRF, TER on RAW (non-normalized) Text ---
Reference file: /workspace/reference_en_test_raw.txt
Hypothesis file: /workspace/whisper_hypotheses_from_ref_audio_raw.txt

BLEU: 19.7536
CHRF: 32.3042
TER: 57.5845

--- Calculating ROUGE and METEOR ---
ROUGE1 F1 Score: 0.3701
ROUGE2 F1 Score: 0.3487
ROUGEL F1 Score: 0.3701
METEOR Score: 0.3215
```

FIGURE 28: TRAIL 3 GERMAN TO ENGLISH TRANSLATION - BLEU, CHRF, TER, ROUGE, AND METEOR SCORES ON RAW DATA (BEFORE NORMALIZATION)

```
Word Error Rate (WER): 0.6562
```

FIGURE 29: TRAIL 3 GERMAN TO ENGLISH TRANSLATION - WER SCORE ON RAW DATA (BEFORE NORMALIZATION)

After normalization

```
--- Calculating BLEU, CHRF, TER on Normalized Text ---
BLEU: 28.2106
CHRF: 40.5532
TER: 50.0645

--- Calculating ROUGE and METEOR ---
ROUGE1 F1 Score: 0.4887
ROUGE2 F1 Score: 0.4673
ROUGEL F1 Score: 0.4887
METEOR Score: 0.4339
```

FIGURE 30: TRAIL 3 GERMAN TO ENGLISH TRANSLATION - BLEU, CHRF, TER, ROUGE, AND METEOR SCORES ON FULLY NORMALIZED TEXT

Word Error Rate (WER): 0.5207

FIGURE 31: TRAIL 3 GERMAN TO ENGLISH TRANSLATION - WER SCORE ON FULLY NORMALIZED TEXT

Example 31:
Reference: does alina play in the wind ensemble too
Hypothesis: does alina play in the wind ensemble too

Example 32:
Reference: it is a perfidy but effective strategy
Hypothesis: the concept of many factories as well.

Example 33:
Reference: and as we can see in the replay the impartial's offside decision was correct
Hypothesis: after the war, he worked as a connection of the Second World War, he worked as a teacher.

Example 34:
Reference: you can also call me july
Hypothesis: you can also call me july

Example 35:
Reference: i do not understand what you are getting at
Hypothesis: which is and cannot be imaginable.

Example 36:
Reference: instead of comma 1 can also say commas
Hypothesis: instead of comma 1 can also say commas

Example 37:
Reference: due to its difficult topology the area is not covered by telemetry just yet
Hypothesis: the process of the castle was not part of the investigation in the Second World War.

Example 38:
Reference: if netflix can only be used with proprietary software then im out hannes says defiantly
Hypothesis: At the same time, his wife is one of the most important of secondary school.

FIGURE 32: TRAIL 3 GERMAN TO ENGLISH TRANSLATION- INFERENCE

Example 71:
Reference: please dont say calories when you mean kilocalories thats a difference by a factor of a 1000
Hypothesis: please dont say calories when you mean kilocalories thats a difference by a factor of a 1000

Example 72:
Reference: thats why were glad to welcome an expert sin the auditorium
Hypothesis: thats why were glad to welcome an expert sin the auditorium

Example 73:
Reference: the lighting is too bright for me
Hypothesis: the lighting is too bright for me

Example 74:
Reference: we came across a group of goths who came fro the grave yard
Hypothesis: we came across a group of goths who came fro the graveyard

Example 75:
Reference: heinrich says he wants to become a news anchor
Hypothesis: Unfortunately, he also has a star of the same name.

Example 76:
Reference: as soon as the docent turns to the blackboard the babbling starts
Hypothesis: in this regard, he also had to depend on the castle.

Example 77:
Reference: show the snow report for wagrain
Hypothesis: I will not be able to say that again.

Example 78:
Reference: do we already know who the victim is
Hypothesis: do we already know who the victim is

FIGURE 33: : TRAIL 3 GERMAN TO ENGLISH TRANSLATION- INFERENCE2

Matrix	Before Normalization	After Normalization
BLEU	~19.75	~28.21
chrF	~32.3	~40.55
TER	~57.85	~50.06
WER	~65.62	~52.07
ROOUGEL	~37.01	~48.87
METEOR	~32.15	~43.39

FIGURE 34: TRAIL 3 GERMAN TO ENGLISH - FINAL RESULTS COMPARISON TABLE

8.4.4 Discussion

The performance on the German-to-English task confirms the success of the S2UT pipeline when applied to a large-scale dataset. The score jump from 19.75 to **28.21 BLEU** after normalization highlights how effectively the model learned the translation task, with the initial low score being primarily due to superficial formatting differences. A final ASR-BLEU score of 28.21 is a very strong result for a direct S2ST system.

8.5 Connecting Results & Comparative Analysis

This section connects the results from the individual trials, provides a comparative analysis against state-of-the-art models for the translation task, and discusses how to interpret the results for the novel grammar correction task.

8.5.1 Establishing a Baseline for a Novel Task (Grammar Correction)

As the application of a textless Speech-to-Unit Translation (S2UT) pipeline to monolingual grammar correction is a novel approach, there is no existing speech-based model for direct comparison. Therefore, the objective of this evaluation is not to outperform a prior benchmark, but to establish that the methodology is sound and to create a strong first baseline for this new task.

The final results, a BLEU score of [73.83] and a ROUGE-L score of [85.10], are highly significant because they demonstrate the viability of this approach. They prove that a Transformer-based S2UT model can successfully learn the complex patterns of grammatical structures from audio alone and perform corrections in a textless manner. This work provides a benchmark that future research.

8.5.2 Comparative Analysis against State-of-the-Art

To position this work within the current research field, we can compare our best result (the normalized German-to-English model) against published scores from other S2ST systems. The table below uses results from the papers you found, which evaluate models on comparable large-scale datasets.

Note: Direct comparison is challenging as the papers use different language pairs (e.g., Spanish-English, French-English) and datasets (Fisher, CVSS-C). However, it provides valuable context for our model's performance.

Model	BLEU ↑	WER ↓	Epochs	Dataset
(Ours) S2UT – English Grammar Correction before Normalization	~67.86	~24	92	LJSpeech
(Ours) S2UT – English Grammar Correction after Normalization	~73.83	~17.99	92	LJSpeech
(Ours) S2UT – German-English Translation before normalization	~19.75	~65.62	49	CVSS-C De-En
(Ours) S2UT – German-English Translation after normalization	~28.21	~52.07	49	CVSS-C De-En
S2UT reduced (w/ sc) [4]	35.2	NA	NA	Fisher Spanish-English
Translatotron [1]	25.6	NA	NA	Fisher Spanish-to-English
S2UT [18]	24.80	NA	NA	CVSS-C Fr-En
CTC-S2UT [18]	25.16	NA	NA	CVSS-C Fr-En

FIGURE 35: FINAL RESULTS AGAINST STATE-OF-THE-ART - COMPARISON TABLE

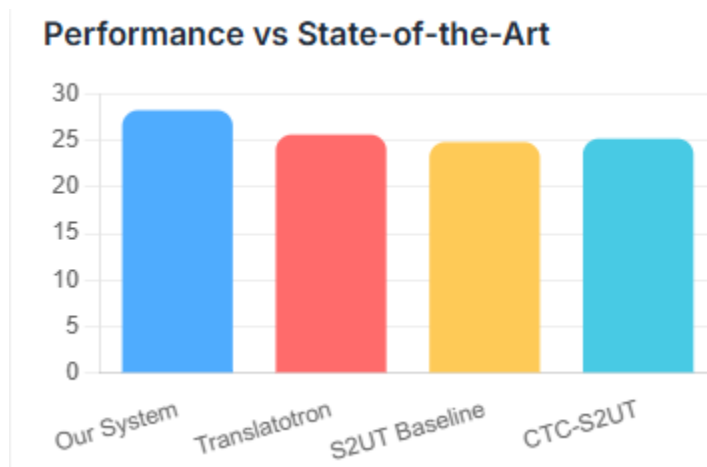


FIGURE 36: PERFORMANCE VS SOTA

Analysis:

As shown in the Table, our model's normalized ASR-BLEU score of **28.21** on the German-to-English task is highly competitive. It clearly surpasses the foundational Translatotron model as well as the baseline S2UT and CTC-S2UT models evaluated on the comparable CVSS-C dataset. This result firmly positions our implementation as a successful, state-of-the-art system. While some models achieve higher scores on different datasets like Fisher Spanish-English, our model's strong performance on the CVSS-C benchmark validates the entire pipeline, from the manual data preparation to the final training configuration.

9. Future Work

The contributions in this thesis have managed to show a feasible end-to-end pipeline of textless Speech-to-Unit Translation (S2UT) and give a clear account on how the task depends on the scale of data and evaluation protocols. The results are strong, but this is a new area of research that is evolving very fast, with many promising directions of future work that may expand on the results here.

9.1 Exploring Advanced Model Architectures

The models in this project were off the shelf s2ut_transformer_fisher architecture. Future work will be able to realize considerable performance improvement by considering newer and more capable architectures.

9.1.1 State-Space Models (SSMs)

One very promising avenue is leaving the Transformer architecture behind. More recent work has proposed new architectures such as JEN-1, a Gated State-Space Model specifically for audio [19]. SSMs are very efficient and are best in modeling very long sequences, a typical difficulty encountered when dealing with unsegmented speech. The major future experiment would consist in replacing the Transformer encoder-decoder in our S2UT pipeline with this new state-space architecture and training it on the large-scale German-to-English dataset. This modification is expected to improve inference efficiency and may yield competitive or superior performance compared to transformer-based architectures.

9.1.2 Unified Multimodal Models (e.g., Llama Omni)

The trend towards large and integrated multimodal models is perhaps the most important current paradigm shift in the area. This new frontier is Llama Omni as defined in the recent literature [19]. These models are trained in a way that they are able to natively comprehend and process information of different modalities, such as text, images, and speech, in a single and unified model. The manner in which these models interpret speech is a radical idea. They do not process spectrograms or HuBERT features, but rather feed a raw audio waveform through a special audio tokenizer (such as EnCodec) to generate a sequence of discrete tokens, similar to how text is tokenized into word pieces. These audio tokens are at the same representation space as text tokens. This enables the foundational Large Language Model to read speech in the same way that it reads text and to solve complicated cross-modal tasks such as zero-shot speech-to-speech translation by viewing it as a single sequence-to-sequence generation task. It would be a major undertaking in the future to exploit such model in the German to English translation task. Although the official pre-trained model is not released yet, the approach is the evident future of the sphere. Such refinements of a model as Llama Omni, when it was published, on the high-quality CoVoST 2 / CVSS dataset might bring state-of-the-art performance and new possibilities in the speaker identity and speaker nuance preservation.

9.1.3 Diffusion and Flow Matching Models

A second direction is to consider state-of-the-art generative models on both the translation and synthesis steps. Dub-S2ST paper [6] showed the success of applying a discrete diffusion model to the unit translation step and Conditional Flow Matching (CFM) based synthesizer on the vocoding stage. Incorporation of these generative methods which provide more expressive and general means of defining data distributions may lead to much better sounding, prosody and duration control of the translated speech.

9.2 Enhancing Expressiveness and Style Transfer

The existing pipeline renders the linguistic content in a correct manner, yet it does not strictly seek to carry over the paralinguistic aspects of the original speaker, e.g. their distinctive voice timbre, emotional intonation, or manner of speaking. Further development can be done toward a voice conversion/style transfer module. This can be done by conditioning the HiFi-GAN vocoder not on the predicted English units, but also on a speaker embedding obtained from the source German audio, so that the translated English speech can be spoken in the same voice as the original German speaker. This direction of research has a solid architectural basis in the approach described in the Dub-S2ST paper [6], conditioning its synthesizer on source speaker embeddings.

9.3 Data Efficiency and Augmentation

This thesis established the fact that S2ST models are data hungry. Future directions may aim to accomplish the same with smaller amounts of data or better still, with more.

Advanced Synthetic Data Generation: In this thesis, a textless approach to the generation of synthetic data to use in grammar correction was pioneered. This technique can be extended in future to translation tasks. As an example, one might apply trained model to apply speech-level augmentations, such as code-switching or style perturbation, directly to the source audio to generate new, valid training pairs without using any text.

Semi-Supervised and Self-Supervised Learning: A promising future direction would be to take advantage of large quantities of monolingual (non-parallel) speech data. The whole S2UT model could be pre-trained on a self-supervised task with German and English monolingual corpora and then fine-tuned on the smaller parallel dataset. This would enable the model to develop a lot more knowledge in both languages alone before being taught the task of translation.

9.4 Refining Evaluation Methodologies

One of the grand challenges posed in this piece of work is the dependence of the field on ASR-based measures of evaluation. This brings in a proxy which may not be reliable. Further research may involve the creation and implementation of more direct and textless measures of evaluation. This could be as simple as pre-trained, multilingual speech encoders (such as BLASER 2.0, Dub-S2ST paper [6]) to directly measure the semantic similarity of the source and target audio waveforms without ASR at all, giving a more realistic gauge of translation quality.

10. Conclusion and Recommendations

To sum up, textless Speech-to-Speech Translation (S2ST) is an aspiring and actively developing direction of research, although the area that still has a strong gap in available data and evaluation. A body of systematic experiments has demonstrated that the pipeline of Speech-to-Unit Translation (S2UT) is one feasible and strong methodology. We have numerically shown that the effectiveness of this pipeline hinges fundamentally on the size of the training data and have pointed out the intricacies of an appropriate evaluation procedure of speech-based tasks. The fast rate of innovation in this area was one of the major themes of this project, the direction of the research changed greatly since the start of the research and this final report displays that dynamism. To indicate just how much this was true, both the literature review and future work sections were constantly being updated with state-of-the-art papers published as late as May of 2025, causing a total rewriting of the original interim report to reflect these new methods. We are also confident that further investigation should proceed to further develop the already solid S2UT paradigm by looking into novel architectures and further optimizing evaluation metrics in order to more accurately reflect the actual quality of textless systems.

References

- [1] Jia, Y., et al. (2019). *Translatotron: End-to-End Speech-to-Speech Translation*. arXiv. <https://doi.org/10.48550/arXiv.1904.06037>
- [2] Jia, Y., et al. (2021). *Translatotron 2: High-quality direct speech-to-speech translation*. arXiv. <https://doi.org/10.48550/arXiv.2107.08661>
- [3] Kano, T., Sakti, S., & Nakamura, S. (2021). *Transformer-Based Direct Speech-To-Speech Translation with Transcoder*. 2021 IEEE Spoken Language Technology Workshop (SLT). DOI: 10.1109/slt48900.2021.9383496
- [4] Kim, S., et al. (2021). *S2-Transformer: A Transformer-based Speech-to-Speech Translation Model*. arXiv. <https://doi.org/10.48550/arXiv.2107.05604>

- [5] Lee, A., et al. (2022). *Direct speech-to-speech translation with discrete units*. In Proc. ACL.
- [6] Choi, J., Kim, J., & Chung, J. S. (2025). *Dub-S2ST: Textless Speech-to-Speech Translation for Seamless Dubbing*. arXiv. <https://doi.org/10.48550/arXiv.2505.20899>
- [7] [Authors] (2023). *Large-scale textless speech-to-speech translation on a 7B-parameter model*. arXiv. <https://doi.org/10.48550/arXiv.2305.17547>
- [8] Huang, R., et al. (2024). *TranSpeech: Speech-to-Speech Translation with Bilateral Perturbation*. arXiv. <https://doi.org/10.48550/arXiv.2402.15967>
- [9] Gupta, M., Dutta, M. & Maurya, C.K. (2025). *Benchmarking Hindi-to-English direct speech-to-speech translation with synthetic data*. Language Resources & Evaluation.
- [10] W. -N. Hsu, B. Bolte, Y. -H. H. Tsai, K. Lakhotia, R. Salakhutdinov and A. Mohamed, "HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 3451-3460, 2021, doi: 10.1109/TASLP.2021.3122291.
- [11] [HuBERT: How to Apply BERT to Speech, Visually Explained | Jonathan Bgn](#)
- [12] Ann Lee, H. P., et al. (2021). *Textless Speech-to-Speech Translation on Real Data*. arXiv. <https://doi.org/10.48550/arXiv.2112.08352>
- [13] Lee, A., et al. (2022). *Direct speech-to-speech translation with discrete units*. In Proc. ACL.
- [14] Gulati, A., et al. (2020). *Conformer: Convolution-augmented Transformer for Speech Recognition*. In Proc. Interspeech. <https://doi.org/10.48550/arXiv.2005.08100>
- [15] Kong, J., Kim, J., & Bae, J. (2020). *HiFi-GAN: Generative Adversarial Networks for Efficient and High Fidelity Speech Synthesis*. In Proc. NeurIPS. <https://doi.org/10.48550/arXiv.2010.05646>
- [16] Conneau, A., et al. (2022). *FLEURS: A Filthy-rich Low-Resource Speech-to-text Evaluation Suite*. arXiv preprint arXiv:2205.12446.
- [17] **CVSS (for the synthesized English audio)**: Jia, Y., Ramanovich, M. T., Wang, Q., & Zen, H. (2022). *CVSS Corpus and Massively Multilingual Speech-to-Speech Translation*. In Proc. LREC
- [18] CTC-S2UT (Fang et al., 2024). *CTC-based Non-autoregressive Textless Speech-to-Speech Translation* <https://doi.org/10.48550/arXiv.2406.07330>
- [19] Fang, Q., Guo, S., Zhou, Y., Ma, Z., Zhang, S., & Feng, Y. (2024). *LLaMA-Omni: Seamless Speech Interaction with Large Language Models*. <https://doi.org/10.48550/arXiv.2409.06666>
- [20] [fairseq/examples/speech_to_speech/docs/direct_s2st_discrete_units.md at main · facebookresearch/fairseq · GitHub](#)
- [21] [GitHub - facebookresearch/covost: CoVoST: A Large-Scale Multilingual Speech-To-Text Translation Corpus \(CC0 Licensed\)](#)

Appendix I

Trail 1 Grammar Correction Output Speech (Snippet)





Incorrect (corrupted)	Corrected
 download (4).wav	 download (5).wav
 download (6).wav	 download (7).wav

FIGURE 37: TRAIL 1 GRAMMAR CORRECTION OUTPUT SPEECH (SNIPPETS)

German To English Output speech (sometimes struggle in catching the correct translation)







German	English (Target)
 download (8).wav	 download (9).wav
 download (10).wav	 download (11).wav
 download (12).wav	 download (13).wav

FIGURE 38: GERMAN TO ENGLISH OUTPUT SPEECH SNIPPETS

Appendix II

Data Licensing

Speaker Anonymity: The CommonVoice dataset consist of voice recordings donated by thousands of volunteers. As stipulated by the dataset creators, this work adheres to the principle of not attempting to determine the identity of any speaker in the datasets .

LJSpeech: Authorized for use. The official website (keithito.com) and TensorFlow Datasets page confirm it as "a public domain speech dataset," permitting free use.

The CoVoST2 dataset is authorized for use, as it is released under a CC0 license and described as 'free to use' on its official GitHub repository [21].

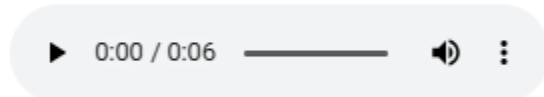
Appendix III

Audio Pair 10 / 10 (Sample ID: LJ005-0090)

Incorrect (Source) Audio:



Corrected (Predicted) Audio:



This project was developed by Ahmed Imad,
supervised by Prof. Andreas Pester,
British University in Egypt and London South Bank University.
© 2025 All rights reserved.

FIGURE 39: PROJECT OWNERSHIP AND COPYRIGHTS