



# Intelligibility of Text-to-Speech Systems for Mathematical Expressions

Sujoy Roychowdhury<sup>1</sup>, Ranjani H.G.<sup>1</sup>, Sumit Soman<sup>1</sup>, Nishtha Paul<sup>2</sup>, Subhadip Bandyopadhyay<sup>1</sup>,  
Siddhanth Iyengar<sup>3</sup>

<sup>1</sup>Ericsson R&D, Bangalore, India; <sup>2</sup>IIT Bangalore, India; <sup>3</sup>BITS Pilani, India

{sujoy.roychowdhury, ranjani.h.g, sumit.soman, subhadip.bandyopadhyay}@ericsson.com

## Abstract

There has been limited evaluation of advanced Text-to-Speech (TTS) models with Mathematical eXpressions (MX) as inputs. In this work, we design experiments to evaluate quality and intelligibility of five TTS models through listening and transcribing tests for various categories of MX. We use two Large Language Models (LLMs) to generate English pronunciation from LaTeX MX as TTS models cannot process LaTeX directly. We use Mean Opinion Score from user ratings and quantify intelligibility through transcription correctness using three metrics. We also compare listener preference of TTS outputs with respect to human expert rendition of same MX. Results establish that output of TTS models for MX is not necessarily intelligible, the gap in intelligibility varies across TTS models and MX category. For most categories, performance of TTS models is significantly worse than that of expert rendition. The effect of choice of LLM is limited. This establishes the need to improve TTS models for MX.

**Index Terms:** TTS, mathematical expressions,  $\LaTeX$ , listening tests, perception experiments, large language models.

## 1. Introduction

There have been significant advances in Text-to-Speech (TTS) models alongside those of Large language models (LLMs) in the recent years [1, 2, 3]. These have resulted in high quality TTS outputs that are nearly indistinguishable from human speech [2, 3]. Typical applications for expressive speech consider and evaluate audio versions of eBooks or podcasts. However, there has been limited (and recent) attention to audio versions of scientific text based content. It typically includes Mathematical eXpressions (MX), tabular data, graphs, or block diagrams, which have higher complexity than conversational or textual content alone. In particular, for TTS MX outputs, it has been established that existing objective TTS evaluation metrics, such as Word Error Rate (WER) with respect to the original transcriptions, does not suffice [4].

It is expected that TTS output (referred to AudioMX, henceforth) for MX input is such that a listener can discern and hence transcribe the MX. To the best of our knowledge, there has been no work which addresses perception and intelligibility aspects of TTS outputs for MX through listening experiments. We design experiments to systematically understand challenges of TTS models in converting MX to AudioMX and their interpretation by human listeners. We present both subjective and objective metrics to evaluate human interpretability captured through transcription in  $\LaTeX$ . We use outputs of five commercially available and popular TTS models (which are considered as closed box or opaque systems, in this work).

### 1.1. Existing work

One of the earliest known MX related datasets is Handwritten and Audio Dataset of Mathematical Expressions (HAMEX) [5], which associated audio and handwritten MX. It was used to consider audio as an additional modality for recognition of handwritten equations [6] and not for the TTS problem. TTS for MX is challenging and earlier works [7, 8, 9, 10] utilized rules and heuristics for converting MX into text of the spoken language (Polish in [8] and Sepedi in [10]), followed by a TTS model for the speech outputs - the TTS has not been evaluated.

*MathSpeech* [11] developed a benchmark dataset and an approach to transcribe MX from audio (extracted from YouTube video lectures) to  $\LaTeX$  using fine-tuned small Language Models; this involved correcting Automatic Speech Recognition (ASR) errors. *MathBridge* [4] introduced a dataset of spoken MX and their corresponding  $\LaTeX$  representations, benchmarked on T5-large [12] using sacreBLEU [13] as a metric. Speech to  $\LaTeX$  challenges are discussed along two approaches - LLMs as post-processing step to ASR, and multimodal LMs for English and Russian languages in [14]. All of these are ASR systems for transcribing MX from speech.

TTS for MX has been proposed and evaluated in *Math-reader* [15] and *MathVision* [16]. The former uses the *MathBridge* dataset to fine-tune T5 model to convert  $\LaTeX$  to Spoken English sentences. However, it assumes TTS models considered are error-free [15]. In *MathVision*, YOLO v7 vision models classify MX to categories, followed by LSTM models to convert output of YOLO models to a sentence in natural language. However, we note that both these works assume TTS systems are error free, and do not evaluate the TTS systems. The need for prosody in TTS outputs for MX inputs has been observed in [17].

In this work, we focus on intelligibility and perceived quality of some of the popular TTS models through listening tests. A major distinction from the existing works and ours is that we focus on evaluating commercial, state-of-the-art (SOTA) TTS models [18] via MOS and LCER. We study the effect of choice of TTS models for various categories of MX. We now present our Research Questions (RQs) and key contributions.

### 1.2. Research Questions and Contributions

The objective of TTS models for MX is that the output audio conveys the content in an intelligible way to the listener. It is critical to be able to follow and comprehend MX correctly. Note that this is not same as cascaded system of TTS with ASR (discussed in Section 5.3). We consider the following RQs.

- **RQ1:** - Can we measure a listener's perception of a TTS audio output of a MX? Does their ability to transcribe it correctly depend on the MX category?

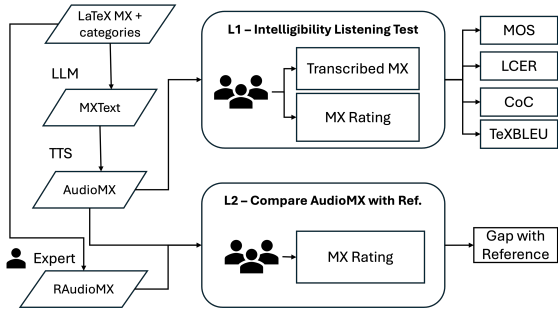


Figure 1: Dataset formats, experiment design and metrics.

- **RQ2**:- Is there any perceived difference between an human expert rendered audio recording of a MX (RAudioMX) and TTS output MX (AudioMX) in terms of intelligibility, quality, clarity and/or speed?

To the best of our knowledge, this is the first study of the TTS models for MX with a focus on user perception and intelligibility. Our **main contributions** are:

- Establish that the intelligibility of TTS outputs varies with the MX category, and that the relative performance of TTS models is dependent on the type of the MX.
- Demonstrate that there is a gap between user’s perception of their understanding of the MX and the ability to transcribe it.
- Identify categories of MX where there is a gap between TTS outputs and reference rendering by experts.
- Qualitatively show that challenges in intelligibility are related to prosody and duration of the synthesis of (parts of) MX.

## 2. Dataset and Methodology

### 2.1. Dataset

There are limited datasets focusing on TTS using MX. HAMEX [19] had handwritten MX and their audio recordings, was one of the earliest, but to the best of our knowledge, is no longer publicly available. The Kaggle dataset Handwritten Mathematical Equations (HME) [5], a superset of HAMEX, has the handwritten MX and  $\LaTeX$  ground truth, but not audio files.

For this work, we consolidate MX from different folders of HME to include categories such as matrices, de-duplicate and obtain a dataset of 3141 MX. We manually categorize these into 8 different categories. Table 1 lists the categories and a sample expression for each category. We sample a total of 120 expressions from the 8 categories for the listening tests.

Though *Mathspeech* [11] (comprising of audio clippings from YouTube videos of classrooms) has recently been released, it is primarily designed for ASR and was released after our research efforts. It also has fewer MX categories than HME.

### 2.2. Methodology

We consider Seamless (SL) [20], an open-source TTS model, and four publicly available models via Application Program Interfaces (APIs) *i.e.*, Google’s Text-to-Speech AI API (GA) [21], Microsoft Azure AI Speech (AZ) [22], OpenAI Text-to-speech (OA) [23], Amazon Polly (AP) [24]. The models we have chosen are among the SOTA TTS models [18].

#### 2.2.1. Direct processing of MX in TTS

We provide the original  $\LaTeX$  expressions to the TTS models - however, none of the models perform well. After a manual evaluation for a few expressions across all models, we did not proceed with this approach.

Category	No. of expressions	QWEN	Sample expression
Advanced Algebraic	16	12	$10x^2\sqrt{5+2\sqrt{5}}+5x-\sqrt{5+2\sqrt{5}}$
Basic Algebraic	16	12	$z-2$
Fractions	16	12	$\frac{1}{\sqrt{t}}$
Calculus	16	12	$\int \cos t dt = \sin t$
Matrices	8	8	$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$
Numerics	16	12	$14+61=75$
Sq. Roots and Trig	16	12	$e^{i\alpha t} \cos \beta t$
Summation	16	12	$\sum_{i=1}^{36} i = 666$
Total	120	92	

Table 1: Categories used in our experiments along with a sample expression. The column QWEN denotes number of pronunciations from QWEN model, rest are from GPT4

### 2.2.2. Use of LLM for pronunciation

To overcome the above challenge (Section 2.2.1), we present the expressions to a LLM with the following prompt - “Provide a pronunciation for the mathematical expression given in Latex below”. We did not explore chain of thoughts (CoT) and other advanced prompting techniques as we focus on evaluation of output of TTS systems for intelligibility. We focus on using an open-source model (QWEN2.5-7b) which is a relatively small LLM. However, in addition, for some of the expressions, we also provide them to GPT4 for comparison. Table 1 shows the count of MX for which we got pronunciation from QWEN per category. The rest are from GPT4. This is an alternative to fine-tuning the T5 model step or LSTM models block in [4, 16] respectively. Use of an existing LLM also reduces training costs. We call the generated pronunciation as MXText.

### 2.2.3. Text to Speech Models

MXText is input to each TTS model in order to obtain the corresponding audio output (referred to as AudioMX). AudioMX are inputs to our listening tests to measure intelligibility and quality.

## 3. Experimental design

Figure 1 shows our experimental design of two listening tests.

### 3.1. L1 - Intelligibility Listening Tests

To address RQ1, the L1 listening test measures intelligibility of MX using only AudioMX. The listeners are not provided with the ground truth MX, the LLM or the TTS model corresponding to the AudioMX. The listeners are asked to transcribe the MX in  $\LaTeX$  to the best of their understanding from the AudioMX. The MX written in  $\LaTeX$  is rendered as preview immediately so that any correction can be made. Listeners can play the audio as many times as required and/or change speed (0.5x to 2x of the actual speed). If unable to comprehend and/or transcribe the AudioMX, they were asked to enter “Did not understand”. The listeners also provide their opinion score on the audio based on intelligibility on a Likert scale of 1 – 5 (1 being the poor and 5 being the excellent). The audio rating is considered to be 1 in case they could not understand the AudioMX. The Mean Opinion Score (MOS) from these ratings provides a subjective view of the perceptual difficulty in audio understanding. The listener transcribed  $\LaTeX$  MX provides an objective view of the intelligibility of the AudioMX.

For L1 test, we sample 120 MX across 8 categories (Table 1). Each listener gets 2 questions from the matrices category and 4 questions from each of the other categories. With 5 TTS models per MX, we evaluate a total of 600 AudioMX. Each AudioMX is rated by 3 listeners - thus we obtain a total of 1800 opinion scores and listener transcribed  $\LaTeX$  MX. The L1 test

Metric Category	Mean Opinion Scores (MOS)					Count of Correct (%)					L <sup>A</sup> T <sub>E</sub> X Character Error Rate					TeXBLEU				
	AP	AZ	GA	OA	SL	AP	AZ	GA	OA	SL	AP	AZ	GA	OA	SL	AP	AZ	GA	OA	SL
Advanced Algebraic	3.9	3.5	<b>4.3</b>	3.9	3.1	33.3	39.6	<b>52.1</b>	<b>52.1</b>	33.3	0.1	0.1	<b>0.0</b>	0.1	0.2	<b>0.9</b>	0.8	<b>0.9</b>	<b>0.9</b>	0.8
Basic Algebraic	<b>4.0</b>	3.9	3.6	3.5	3.2	<b>70.8</b>	<b>70.8</b>	47.9	66.7	33.3	<b>0.1</b>	<b>0.1</b>	0.2	0.2	<b>0.1</b>	<b>0.9</b>	<b>0.9</b>	0.8	0.8	0.8
Calculus	3.6	3.6	3.7	<b>3.9</b>	2.6	54.2	50.0	39.6	<b>60.4</b>	18.8	<b>0.1</b>	<b>0.1</b>	<b>0.1</b>	<b>0.1</b>	0.3	<b>0.8</b>	<b>0.8</b>	0.7	<b>0.8</b>	<b>0.6</b>
Fractions	4.1	<b>4.3</b>	4.2	4.1	3.9	58.3	58.3	<b>66.7</b>	62.5	47.9	0.1	<b>0.0</b>	0.1	<b>0.0</b>	0.1	<b>0.8</b>	<b>0.8</b>	<b>0.8</b>	<b>0.8</b>	<b>0.8</b>
Matrices	3.1	3.1	3.1	3.5	<b>3.6</b>	20.8	16.7	25.0	<b>29.2</b>	20.8	0.2	0.2	0.2	<b>0.1</b>	<b>0.1</b>	0.7	0.7	0.7	<b>0.8</b>	<b>0.8</b>
Numerics	<b>4.9</b>	4.8	<b>4.9</b>	<b>4.9</b>	4.7	97.9	97.9	97.9	<b>100.0</b>	<b>100.0</b>	<b>0.0</b>	<b>0.0</b>	<b>0.0</b>	<b>0.0</b>	<b>0.0</b>	<b>1.0</b>	<b>1.0</b>	<b>1.0</b>	<b>1.0</b>	<b>1.0</b>
Sq. Roots and Trig	3.6	4.0	<b>4.4</b>	4.1	3.3	50.0	54.2	<b>64.6</b>	58.3	29.2	0.1	0.1	<b>0.0</b>	0.1	0.1	<b>0.9</b>	0.8	<b>0.9</b>	<b>0.9</b>	0.8
Summation	3.0	3.3	<b>3.5</b>	<b>3.5</b>	2.3	41.7	<b>58.3</b>	45.8	45.8	39.6	0.2	<b>0.1</b>	0.2	0.2	0.4	0.7	<b>0.8</b>	0.7	0.7	0.5

(a) MOS

(b) CoC as %

(c) LCER

(d) TeXBLEU

Table 2: Average MOS, LCER, CoC (in %) and TeXBLEU metrics across categories and TTS models for L1-listening test. Highlighted numbers indicate best average scores across models.

was conducted in 60 batches, each having 30 AudioMX. We enrolled 49 listeners for this, a few of who did at most 3 batches. All of the listeners who participated in the listening tests have age range between 18-50 years, a background in science or engineering education and are familiar with L<sup>A</sup>T<sub>E</sub>X.

### 3.2. Evaluation of transcribed MX

Based on challenges in evaluating MX [25, 4], we consider the following three approaches to compare transcribed MX with the ground truth MX. The results are shown in Table 2.

- Four evaluators manually compare the ground truth MX and transcribed MX marking them as correct or incorrect. The L<sup>A</sup>T<sub>E</sub>X was rendered for ease of correction. Each MX was marked correct or wrong based on the equivalence of the relations. The evaluators do not have of any details of the response (such as the MX category, LLM, TTS model, listener identity who had transcribed the MX, or the original audio rating). We call the metric derived out of this expert validation as Count-of-Correct (CoC). We do not increment the counter in case the response is “*Did not understand*”.
- For evaluating how close the transcribed MX is to the original MX we use LCER. This is calculated by removing spaces from the L<sup>A</sup>T<sub>E</sub>X representation and then computing the Levenshtein distance [26]. This is then normalized by min-max normalization to transform it in the range of [0, 1]. In case the response is “*Did not understand*”, we set LCER to 1.
- We compute TeXBLEU [25] to evaluate similarity between the ground truth and user-written MX. TeXBLEU captures structural and syntactic differences, providing a score between 0 and 1, higher values indicate greater similarity. For response “*Did not understand*”, the score was set to 0.

Category	AP	AZ	GA	OA	SL
Advanced Algebraic	28.1	34.1	<b>24.4</b>	24.7	51.4
Basic Algebraic	26.7	<b>24.1</b>	36.9	25.6	47.3
Calculus	30.9	<b>9.5</b>	24.1	17.7	66.9
Fractions	4.6	<b>2.7</b>	4.7	6.9	23.2
Numerics	3.0	<b>1.45</b>	2.8	2.7	6.2
Sq. Roots and Trig	33.5	49.5	<b>17.1</b>	21.9	56.9
Summation	42.1	39.9	<b>35.5</b>	36.9	72.5

Table 3: Average of difference of reference score with TTS score for different categories in L2 listening test (lower the better).

### 3.3. L2 - Comparison of TTS AudioMX with Reference

To address RQ2, the L2 listening test is designed to measure user preference of different TTS models along with a hidden reference AudioMX. This provides an indication of the gap with SOTA models for MX. 4 experts are asked to record the MX in a way they would typically communicate to an audience via

audio alone. This is considered as Reference AudioMX (RAudioMX).

The L2 comparative tests are motivated by Multi Stimulus test with Hidden Reference and Anchor (MUSHRA) recommendations [27]. Instead of explicit reference audio, we render the MX visually<sup>1</sup>. Thus for every MX, listeners are asked to rate the TTS AudioMX and (hidden) RAudioMX on a scale of [0,100] with respect to clarity and intelligibility of the displayed MX. The details of the TTS model or LLM corresponding to the AudioMX are not provided. Each MX is rated by 3 listeners. We sample 35 MX from 5 categories (excluding matrices). We split the MX into 15 batches with each listener rating 7 MX (one from each category). Each MX comparison has 6 audio files, including the hidden reference, resulting in 42 score preferences. L2 test is completed by 13 individuals with 2 individuals scoring at most 2 batches each.

## 4. Results and Analysis

Table 2 tabulates the average MOS, LCER, CoC and TeXBLEU metrics (detailed in Section 3.1 and 3.2) across categories and the considered TTS models for L1 listening test. These metrics address RQ1. From Table 2(a)-(d), we observe that there is no TTS model which scores consistently high across categories and across metrics. It is also interesting to note that models having better CoC, lower LCER and high TeXBLEU score imply consistent better MOS scores and vice-versa. Hence to identify the significant main and interactions effects among the factors ‘LLM’, ‘TTS Model’ and ‘Category’ from the L1 scores, we perform An Analysis of Variance (ANoVA) [28] tests on LCER scores where ‘TTS Model’ and ‘Category’ factors are found to be statistically significant ( $p$  value  $< 0.05$ ). We infer that comprehension of MX from a TTS audio depends on category of MX apart from the TTS model itself. Although ANoVA with MOS as the response additionally identifies factor ‘LLM’ also as significant, we conclude that the listener rating of AudioMX is not reflected in their MX transcription. Further, we manually check the correctness of pronunciations (87.5% overall, 85.9% for QWEN, 92.9% for GPT4) and perform ANoVA with correctly pronounced MX. We observe that ‘LLM’ becomes an insignificant factor with MOS as the response, but there is no change of inference in the study with LCER as the response. Thus, when the pronunciation is correct, LCER and MOS are statistically equivalent across both LLMs, irrespective of the LLM size.

We also analyse CoC considering length of the audio; we observe that audio which are relatively long are more difficult to transcribe. The CoC was at 64% for AudioMX having duration  $< 3s$ , this reduces to 57.1% considering AudioMX of upto

<sup>1</sup>Screenshot of the listening test UI in supplementary material

Equation	LLM Pronunciation	Expected
$\int_0^{\frac{\pi}{2}} \frac{\sin^2 x}{x^2} dx =$	Integral from 0 to infinity of (sin x squared over x squared) equals pi over 2	Integral from zero to infinity of, sin squared x, over x squared, equals pi over 2
$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	Pronounced as "one by one identity matrix" or simply "identity matrix."	Two by two identity matrix
$2^n - i$	Two to the n minus i power	Two to the n, minus i

Table 4: Representative pronunciation errors (marked in red) identified by an expert and expected pronunciation.

6s duration and 55.0% those upto 12s duration. 97.5% of AudioMX are less than 12s duration, with the longest being 18.2s.

For RQ2, we use the human reference audios - they have average CoC is 97% across categories, average LCER is 0.01 with only one of 35 equations being incorrectly transcribed. From L2 listening test scores (detailed in Section 3.3), we calculate the difference between the listener score for RAudioMX and every TTS AudioMX for each MX. The mean of these per category are presented in Table 3. For all MX, the score for (hidden) RAudioMX is highest. We observe the scores to vary across categories in the results for this test too. Our results show that for MX category 'Numerics', all the models are scored close to reference; for 'Fractions', all models except SL are close to reference. In all other categories, there exists large gaps with respect to the reference audio. The gap observed in Table 3 implies that public API or open source general purpose TTS models do not include the complexities of prosody particular to MX and is in line with discussions in [17].

In conclusion, for RQ1, we observe that outputs for MX of SOTA TTS models are not necessarily intelligible to a human listener. The gap in intelligibility varies across categories and models. For RQ2, we establish that for most categories there is a large gap with respect to the reference indicating a significant scope for improvement in TTS for MX even for SOTA models.

## 5. Discussion

### 5.1. Qualitative Analysis of MX Transcription Error

Some of the patterns and errors observed during process of manual verification of transcribed MX (discussed in Section 3.2), common feedbacks resulting from a focus-group discussion with listeners are detailed below.

- Pronunciation related
  - Confusion among variables such as ('b', 'p'), similar words such as ( $\pi$ , five,  $\phi$ ) has resulted in higher LCER, reduced CoC and TeXBLEU metrics.
  - Variations of TTS outputs of 'cos t' as [koh.sam ti], [kohs ti]; incorrect pronunciation TTS outputs of 'sin t' as [sm ti] instead of [sam ti] has resulted in reduced MOS Scores.
- Prosody related
  - For many MX there was a lack of clarity on the parenthesis or order of operations implied in the MX. For example, the MX  $\frac{a^2-b^2}{a}$  might be transcribed as  $a^2 - \frac{b^2}{a}$ ; similarly  $50 \pm \frac{51}{144}$  with some transcribing it as  $\frac{50 \pm 51}{144}$ . These observations are in line with those in [14].
  - Varied pauses (across TTS models) between words 'square' and 'root' can result in MX  $(-a - b\sqrt{2})$  to be comprehended and transcribed as  $(-a - b^2\sqrt{2})$ . These

increase LCER and reduce CoC and TeXBLEU metrics.

- The TTS models do not consider appropriate pauses nor adjust the speed of audio for more complex MX (such as calculus, summation, or those having equality signs). This results in multiple listenings per MX, causing listening fatigue. The incorrect speed and pause results in perception that AudioMX is unclear and sometimes, rushed; for example, there was no pause between 'integral from zero to infinity' and the variables within integral; these are typical suprasegmental prosodic components of RAudioMX.
- Changing the AudioMX speed was possible but it did not necessarily result in better comprehension of MX.
- Listeners found some of the AudioMX to be longer in duration and hence more challenges in transcribing. This is in line with the duration analysis in Section 4.
- Comprehending AudioMX involving matrices and summation was particularly difficult and many listeners marked them as 'Did not understand'.

### 5.2. Differences with reference

As mentioned in Section 3, RAudioMX is recorded without using LLM-generated pronunciation. Most experts used descriptors like 'whole divided by' for parenthesis e.g. 'a-squared minus b-squared whole divided by a' for  $\frac{a^2-b^2}{a}$ . In addition, appropriate pauses between operands, variables, equality signs are introduced. The scores from L2 listening test indicate that these greatly affect the scoring of an AudioMX to the listener. Prosody for MX therefore needs comprehension of the expression; this is absent across all TTS models except for the simplest of categories such as numerics and fractions.

### 5.3. Automatic Speech Recognition Check

The AudioMX from TTS is passed through ASR model (we choose Whisper-base model [29]). We then compute ASR metrics from this cascaded system - Character Error Rate (CER), Bilingual Evaluation Understudy (BLEU) and Recall-Oriented Understudy for Gisting Evaluation (ROUGE) scores [30] w.r.t. MXText. Average CER is 0.10, BLEU is 0.66 and ROUGE is 0.84 indicating an faithful pipeline, averaged across all TTS models. The correlation of the ASR based metrics (CER, BLEU and ROUGE) with the transcription based metrics (CoC, LCER, TeXBLEU) is very low - the absolute correlation value between any ASR metric and any transcription based metrics is less than 0.15. Thus, whilst the TTS-ASR cascaded models can result in high overall ASR metrics w.r.t. MXText, they are not necessarily intelligible to a human listener.

## 6. Conclusions & Future Work

Our research establishes the variability of TTS performance (for RQ1) based on category and the gap with human rendered audio (for RQ2). We observe that LLMs, such as GPT4, can provide the pronunciations from MX. Any real-world system will need an analysis of the categories of expressions relevant in that application; the choice of TTS models will be based on relative performance for these categories. One may need custom-trained / fine-tuned TTS models because of lack of prosody for MX in the existing publicly available TTS models.

Future work involves fine-tuning models to improve performance on audio rendering of MX and evaluating them with current baselines.

## 7. References

- [1] W. Cui, D. Yu, X. Jiao, Z. Meng, G. Zhang, Q. Wang, Y. Guo, and I. King, "Recent advances in speech language models: A survey," *arXiv preprint arXiv:2410.03751*, 2024.
- [2] H. Barakat, O. Turk, and C. Demiroglu, "Deep learning-based expressive speech synthesis: a systematic review of approaches, challenges, and resources," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2024, no. 1, p. 11, 2024.
- [3] S. E. Eskimez, X. Wang, M. Thakker, C. Li, C.-H. Tsai, Z. Xiao, H. Yang, Z. Zhu, M. Tang, X. Tan, Y. Liu, S. Zhao, and N. Kanda, "E2 tts: Embarrassingly easy fully non-autoregressive zero-shot tts," pp. 682–689, 2024.
- [4] K. Jung, S. Hyeon, J. Y. Kwon, N.-J. Kim, H. G. Ryu, H.-J. Lee, and J. Do, "Mathbridge: A large corpus dataset for translating spoken mathematical expressions into latex formulas for improved readability," *CoRR*, vol. abs/2408.07081, 2024. [Online]. Available: <https://doi.org/10.48550/arXiv.2408.07081>
- [5] R. Tatman, "Handwritten mathematical expressions dataset," <https://www.kaggle.com/datasets/rtatman/handwritten-mathematical-expressions/data>, 2017.
- [6] S. Medjkoune, H. Mouchère, S. Petitrenaud, and C. Viard-Gaudin, "Combining speech and handwriting modalities for mathematical expression recognition," *IEEE Transactions on Human-Machine Systems*, vol. 47, no. 2, pp. 259–272, 2017.
- [7] P. Caky, J. Boron, M. Klimo, and K. Bachrata, "Mathematical formulas in text to speech system," *Scientifi Letters of the University of Zilina*, no. 3, 2009.
- [8] A. Bier and Z. Sroczynski, "Adaptive math-to-speech interface," in *Proceedings of the Multimedia, Interaction, Design and Innovation*, 2015, pp. 1–9.
- [9] A. Bier and Z. Sroczynski, "Rule-based intelligent system verbalizing mathematical notation," *Multimedia Tools and Applications*, vol. 78, pp. 28 089–28 110, 2019.
- [10] M. M. Mogale, T. J. Sefara, and T. B. Mokgonyane, "Grammar-driven text-to-speech application for articulation of mathematical expressions," Southern Africa Telecommunication Networks and Applications Conference (SATNAC), 2020.
- [11] S. Hyeon, K. Jung, J. Won, N.-J. Kim, H. G. Ryu, H.-J. Lee, and J. Do, "Mathspeech: Leveraging small lms for accurate conversion in mathematical speech-to-formula," *arXiv preprint arXiv:2412.15655*, 2024.
- [12] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020. [Online]. Available: <http://jmlr.org/papers/v21/20-074.html>
- [13] M. Post, "A call for clarity in reporting BLEU scores," in *Proceedings of the Third Conference on Machine Translation: Research Papers*, O. Bojar, R. Chatterjee, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, C. Monz, M. Negri, A. N ev ol, M. Neves, M. Post, L. Specia, M. Turchi, and K. Verspoor, Eds. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018, pp. 186–191. [Online]. Available: <https://aclanthology.org/W18-6319/>
- [14] D. Korzh, A. Iudin, N. Kuzmin, E. Karimov, M. Skripkin, O. Rogov, and I. Oseledets, "Listening to formulas: Pioneering models and datasets for converting speech to latex equations," 2024. [Online]. Available: <https://openreview.net/forum?id=pflsJ6V6CL>
- [15] S. Hyeon, K. Jung, N.-J. Kim, H. G. Ryu, and J. Do, "Mathreader : Text-to-speech for mathematical documents," 2025. [Online]. Available: <https://arxiv.org/abs/2501.07088>
- [16] M. Awais, T. Ahmed, M. Aslam, A. Rehman, F. S. Alamri, S. A. Bahaj, and T. Saba, "Mathvision: An accessible intelligent agent for visually impaired people to understand mathematical equations," *IEEE Access*, 2024.
- [17] A. Souza and D. Freitas, "Towards a prosodic model for synthesized speech of mathematical expressions in mathml," in *Proceedings of the 9th International Conference on Software Development and Technologies for Enhancing Accessibility and Fighting Info-Exclusion*, 2020, pp. 105–110.
- [18] Labelbox, "Evaluating leading text-to-speech models," 2024, accessed: 2025-02-11. [Online]. Available: <https://labelbox.com/guides/evaluating-leading-text-to-speech-models/>
- [19] S. Quiniou, H. Mouch ere, C. Viard-Gaudin, E. Morin, S. Petitrenaud, S. Medjkoune *et al.*, "Hamex-a handwritten and audio dataset of mathematical expressions," in *2011 International Conference on Document Analysis and Recognition*. IEEE, 2011, pp. 452–456.
- [20] S. et al, "Seamless: Multilingual expressive and streaming speech translation," 2023.
- [21] G. Cloud, "Google text-to-speech," <https://cloud.google.com/text-to-speech>, 2023, accessed: 2025-02-07.
- [22] M. Azure, "Azure ai speech," <https://azure.microsoft.com/en-us/products/ai-services/ai-speech>, 2023, accessed: 2025-02-07.
- [23] OpenAI, "Openai text-to-speech guide," <https://platform.openai.com/docs/guides/text-to-speech>, 2023, accessed: 2025-02-07.
- [24] A. W. Services, "Amazon polly," <https://aws.amazon.com/polly/>, 2023, accessed: 2025-02-07.
- [25] K. Jung, N.-J. Kim, H. Ryu, S. Hyeon, S.-j. Lee, and H.-j. Lee, "Texbleu: Automatic metric for evaluate latex format," *arXiv preprint arXiv:2409.06639*, 2024.
- [26] K. Kukich, "Techniques for automatically correcting words in text," *ACM computing surveys (CSUR)*, vol. 24, no. 4, pp. 377–439, 1992.
- [27] International Telecommunications Union, "ITU-R BS.1534: Method for the Subjective Assessment of Intermediate Quality Levels of Coding Systems," International Telecommunications Union, Tech. Rep. BS.1534, July 2014.
- [28] L. St, S. Wold *et al.*, "Analysis of variance (anova)," *Chemometrics and intelligent laboratory systems*, vol. 6, no. 4, pp. 259–272, 1989.
- [29] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.
- [30] M. Labied, A. Belangour, and M. Banane, "Assessing speech-to-text translation quality: An overview of key metrics," in *2024 International Conference on Decision Aid Sciences and Applications (DASA)*. IEEE, 2024, pp. 1–6.