Evolution of Performance Metrics for Accurate Evaluation of Speech-to-Speech Translation Models: A Literature Review

Gabriel O. Sobola, Emmanuel Adetiba, Olabode Idowu-Bismark, Oluseyi O. Ajayi, Raymond Jules Kala, Abdul taofeek Abayomi, Oluwadamilola Oshin & Olutoyin Olaitan

Engineering Covenant University

ABSTRACT

The translation of speech from a source to speech in a target language with generative artificial intelligence is an area of research that is presently being actively explored. This is aimed at solving global language barriers thereby ensuring seamless communication between the individuals involved. It has been well developed for high-resourced languages like English, Spanish, French and Chinese. Currently, objective evaluation metrics such as Bilingual Evaluation Understudy Scores (BLEUS), and subjective metrics such as Mean Opinion Score Naturalness (MOSN) and Mean Opinion Score Similarity (MOSS) are being used to evaluate the performance of the output of speech-to-speech models. However, low resourced languages are still undeveloped in the area of speech processing applications, especially the African indigenous languages. The output speech in the target language needs to be evaluated to determine the closeness to the ground truth, as well as how natural and intelligible it is to the intended listeners.

Keywords: BERTscore, bilingual evaluation understudy scores (BLEUS), BLASER, leaderboards, mean opinion score naturalness (MOSN), mean opinion score similarity (MOSS), recall oriented understudy for gisting evaluation longest common subsequence (ROUGE-L), speech-to-speech metrics, word error rate (WER).

Classification: LCC Code: P418.02028

Language: English

Great Britain Journals Press

LJP Copyright ID: 392945
Print ISSN: 2631-8474
Online ISSN: 2631-8482

London Journal of Engineering Research

Volume 25 | Issue 4 | Compilation 1.0

Evolution of Performance Metrics for Accurate Evaluation of Speech-to-Speech Translation Models: A Literature Review

Gabriel O. Sobola[α], Emmanuel Adetiba[σ], Olabode Idowu-Bismark[ρ], Oluseyi O. Ajayi[ↀ], Raymond Jules Kala[¥], Abdul taofeek Abayomi[§], Oluwadamilola Oshin[χ] & Olutoyin Olaitan[v]

ABSTRACT

The translation of speech from a source to speech in a target language with generative artificial intelligence is an area of research that is presently being actively explored. This is aimed at solving global language barriers thereby ensuring seamless communication between the individuals involved. It has been well developed for high-resourced languages like English, Spanish, French and Chinese. Currently, objective evaluation metrics such as Bilingual Evaluation Understudy Scores (BLEUS), and subjective metrics such as Mean Opinion Score Naturalness (MOSN) and Mean Opinion Score Similarity (MOSS) are being used to evaluate the performance of the output of speech-to-speech models. However, low resourced languages are still undeveloped in the area of speech processing applications, especially the African indigenous languages. The output speech in the target language needs to be evaluated to determine the closeness to the ground truth, as well as how natural and intelligible it is to the intended listeners. This paper presents a review of trends from the current metrics to emerging ones such as Recall Oriented Understudy for Gisting Evaluation-L (ROUGE-L) and BLASER. The applications of speech models' metrics on various leaderboards and modern AI platforms were also discussed. The outcome shows that while BLEU score and MOSN metrics are prevalent for speech models, there is a need to explore metrics such as ROUGE-L, and BERTScore which are machine translation metric due to their benefits.

Keywords: BERTscore, bilingual evaluation understudy scores (BLEUS), BLASER, leaderboards, mean opinion score naturalness (MOSN), mean opinion score similarity (MOSS), recall oriented understudy for gisting evaluation longest common subsequence (ROUGE-L), speech-to-speech metrics, word error rate (WER).

Author α σ ρ χ: Covenant Applied Informatics and Communications Africa Center of Excellence (CApIC-ACE) & Department of Electrical and Information Engineering Covenant University, Ota 112104, Nigeria.
σ: HRA, Institute for Systems Science, Durban University of Technology, Durban 4001, South Africa.
ↀ: Department of Mechanical Engineering, Covenant University, Ota 112104, Nigeria.
¥: International University of Grand-Bassam, Grand- Bassam, Côte d'Ivoire.
§: Faculty of Engineering, Built Environment and Information Technology, Walter Sisulu University, Buffalo City Campus, East London 5200, South Africa & Department of Information Communication, Summit University, PMB 4412, Offa 250101, Kwara State, Nigeria.
v: Department of Applied Informatics and Mathematical Sciences, Walter Sisulu University, 27, Nelson Mandela Drive, Mthatha 5100, South Africa.

I. INTRODUCTION

The translation of speech from one language, *the source*, to another language, *the target*, demands an efficient evaluation metric for its evaluation. Researchers in the area of speech processing applications are considering objective evaluation metrics such as the Word Error Rate (WER), Bilingual Evaluation

Understudy (BLEU) scores, and BERTScore as well as subjective evaluation metrics such as the Mean Opinion Score (MOS) Naturalness and Similarity for evaluating the output of such models. Currently, there are no standard objective evaluation metrics applied directly to the generated output target speech [1]. This is because all objective evaluation demands speech output to be converted to texts. Hence, there are issues associated with such metrics. For instance, to utilise the BLEU score metric, the output speech needs to be transcribed to texts after which the texts are being compared or evaluated against the reference or ground truth texts. Researchers have observed overtime errors introduced in the obtained BLEU score due to the Automatic Speech Recognition (ASR) model utilised for such task. According to them, using an evaluation metric that takes the output target speech directly will be better than the ASR for computation of BLEU score. The ASR models have inherent errors that alter the expected metric result obtained using the BLEU score. To overcome this issue, some researchers have utilised large ASR models such as Whisper, based on large hours of data training to generate the transcripts to be compared with the ground truth [2 - 5]. Other researchers have also explored MOS Naturalness and MOS Similarity which are both subjective evaluation in conjunction with objective metrics. Here, raters are sourced to evaluate the performance of the generated output speech using their natural instinct to rate the speech model's output. Such metrics are being utilised to evaluate the fluency, accuracy, quality, and correctness of the generated speech, and the ratings are based on human judgments. To further enhance the objective evaluation metrics, it was confirmed in [1] that character-based performance metrics like character-based F1 score (chrF), and character-based BLEU score (chrBLEU) [6] are more robust metrics for speech-to-speech translation and speech synthesis tasks [1]. They were discovered to show a high correlation with human judgment than BLEU and MCD [1]. Some state-of-the-art speech-to-speech translation models such as Whisper [26], Translatotron, Translatotron2, SeamlessM4T and AudioPalm, and speech models developed by other researchers have utilised the aforementioned metrics to evaluate their models.

Metrics such as the Metric for the Evaluation of Translation with Explicit Ordering (METEOR) [1] that shows high correlation with human judgment [7], ROUGE – L, and BLASER have found low utilisation for computation of speech model evaluation. METEOR, ROUGE - L are mostly utilised for texts summarisations, and other Natural Language Processing (NLP) tasks such as machine translation and question answering [7]. Translation Error Rate (TER) is another metric for machine translation [7] which could be explored for speech-to-speech translation tasks [8] as it could be used to evaluate the texts equivalents of the target speech. As a result, it is suggested that researchers could explore ROUGE-L for the speech-to-speech translation tasks since it is evaluated on the texts obtained from generated target speech. The ROUGE – L is a metric that makes use of non-contiguous subsequence obtained from both the predicted texts and ground truth texts by comparing the two texts. It has been shown to have relationship with the well-known statistical metrics such as precision, recall, and f1-scores, by setting the beta parameter in its formula. Other variants of ROUGE such as the ROUGE-S, and ROUGE-W (ROUGE-Weighted) [9] have performed successfully in machine translation tasks [7, 9]. Other common ROUGE used in translation tasks are ROUGE-WE (ROUGE-Word Embedding), ROUGE-G (ROUGE-Graph based) and ROUGE-1, ROUGE-2 [7]. WER, an objective evaluation is mostly utilised for Automatic Speech Recognition (ASR) where the transcribed texts are compared with the reference texts. In addition to this is the BERTScore that is mostly utilized for machine translation models. This has been confirmed to perform better than ROUGE-L, METEOR, and BLEU score due to its high similarity measure between the candidate (machine translated output) and reference or ground truth samples using the cosine similarity procedure [7]. However, it lacks word ordering. It can also be utilized for speech-to-speech translation model by using the transcribed texts rather than the target speech obtained. There is also Cross Lingual Semantic Textual Similarity (XSTS) [10-11], a human metric that measures semantic similarity between a source speech and target translation. At present, more AI platforms and leaderboards are being engaged in the areas of the Natural Language Processing

(NLP) to rate and compare AI models. Some are known to speech models to evaluate the performances of various SOTA models. Examples of such are the Hugging Face Leaderboards, IWSLT Challenge etc. [45-50] and some are specifically for speech-to-speech translation models. Each platform has its own metrics used to evaluate speech models. Going by the aforementioned, there are numerous metrics/models that have been used in the past and are currently being engaged with varying performance levels. Hence, there is a need to evaluate the performance of some of the models to ascertain their efficiency. This is the basis for this study. Various metrics that are used in speech-to-speech translation tasks and speech processing applications were reviewed with a view to comparatively determining their efficiencies.

II. PERFORMANCE METRICS

The evaluation of speech-to-speech translation models are either carried out on the target output speech (Subjective evaluation), transcribed texts of the target output speech using an ASR model (objective evaluation), or spectral representations of the target output speech. At other times, the evaluation could be carried out in subjective approach in terms of comparing the target output speech with the source speech as in the case of Cross Lingual Semantic Textual Similarity (XSTS). Others are statistical evaluation metrics such as Accuracy, Precision, Recall, and F1-Score.

2.1 Subjective Performance Metrics

In this evaluation metrics, raters are hired to judge the output speech obtained from the speech-to-speech model. In order words the evaluation is performed directly on the speech obtained. Examples of such metrics are MOS Naturalness, MOS Similarity, XSTS etc.

2.1.1 Mean Opinion Score (MOS)

This is a subjective performance metric that is based on the judgment made by the observer on the output translated speech of the target language. It is the most utilised for speech- to-speech evaluation metric [3]. It could be MOS Naturalness or MOS Similarity. In this type of subjective evaluation, raters are sought, who then score the output generated speech on a scale from 1 to 5 which could be 1 = Poor, 2 = Satisfactory, 3 = Good, 4 = Very good, and 5 = Excellent.

2.1.1.1 Mean Opinion Score (MOS) Naturalness

In MOS Naturalness, the raters judge the quality, naturalness, and appropriateness of pro nunciation of the speech output on a scale of 1 to 5. In this case, an incorrectly translated natural target output speech will be rated higher [2] when compared with a correctly unnatural target output speech.

2.1.1.2 Mean Opinion Score Similarity

In MOS Similarity, the raters score the output speech obtained by comparing it with a reference or ground truth output (which can be human-generated speech or synthesised) on a scale of 1 to 5 using quality such as the fluency (flow or correctness of grammar), and adequacy (how deviated speech is from its intended meaning or deviation from the ground or reference speech) of the generated output speech.

2.1.2 Cross Lingual Semantic Textual Similarity (XSTS)

This is a subjective evaluation that is carried out by human raters to assess the quality of the translated target speech. It is conducted by comparing the adequacy (how close it is to its intended meaning) of the generated speech to the source speech. This implies that the human rater or judge must be bilingual

to be able to access such target speech for its meaning. The human annotator judges the semantic similarity rather than fluency between the source and target speech [10-11]. Using a score on a scale of 1 to 5, the annotator assigned each language pair (source-translated target speech) for semantic meaning where a score of 3 or more indicates the two speeches are close in terms of meaning being conveyed. XSTS is a subjective evaluation metric that also checks for the audio quality as it is utilised directly on the audio generated output target speech. It was originally developed for texts evaluation [12]. To obtain the final XSTS results, an average value is computed across selected XSTS computed scores by the human annotators.

2.1.3 Blaser

This is a modality agnostic evaluation metric that works on both speech and texts [13-14]. A version of BLASER, BLASER 2.0 utilised in [14] was a modification on the first version [13]. For speech-to-speech translation tasks, it offers the advantage of being text-free unlike the ASR BLEU performance metric. For BLASER 2.0, the source speech, the translated target speech and the reference texts are converted into Sentence-level multimOdal and language-Agnostic Representations (SONAR) embedding vectors (h_{src}, h_{mt}, $and\ h_{ref}$). These vectors are then fed into a small dense neural network for prediction of XSTS scores for each output of the translation for the supervised version of BLASER 2.0. For the unsupervised version, the cosine similarities between the source and target translated output is obtained for BLASER computation.

2.2 Objective Performance Metric

In the objective, the target output speech of the speech-to-speech translation model is fed to an ASR model to obtain the texts or transcripts equivalent of the speech. Evaluation is then performed on these texts to assess the performance of the model. Examples are the WER, BLEU Score, ROUGE-L, BERTScore, etc.

2.2.1 Bilingual Evaluation Understudy (BLEU)

This is an objective n-gram evaluation that involves the comparison of the speech target output with that of the reference or ground truth. To compute the BLEU score, the output of the speech translation is fed to an ASR model to generate the text equivalent. The texts generated are then compared with the ground truth texts, and the BLEU score is computed. Mathematically, the BLEU score is computed using equation (1) [15]:

$$BLEU\ score\ =\ BP\ *\ \exp exp\ \sum_{i=1}^{N} w_i p_i \tag{1}$$

where:

$$BP\ =\ Brevity\ Penalty\ =\ \exp exp\left(1\ -\ \frac{r}{c}\right) \tag{2}$$

which is also computed as:

$$BP\ =\ Brevity\ Penalty\ =\left(1,\ \frac{r}{c}\right) \tag{3}$$

$r\ =\ length\ of\ machine\ translated\ output\ (text\ or\ speech)$
$c\ =\ length\ of\ reference\ translation\ (text\ or\ speech)$
$p_i = n\ -\ \ gram\ modified\ precision\ score\ of\ order\ i,\ which\ is\ given\ in\ equation\ (4)\ as$:

$$p_i = \frac{Count\ Clip\ (matches, max{-}reference{-}count)}{candidatae\ n{-}gram} \tag{4}$$

$N = maximum\ number\ of\ n - gram\ order\ to\ consider\ (usually\ up\ to\ 4)$
$w_i = weight\ for\ n - gram\ precision\ of\ order$

2.2.1.1 Evaluation of BLEU Score Computation

Given the information below for both the machine-translated output (obtained using ASR) and reference output text:

Machine Translation (MT): The picture the picture by me.
Reference (Ref) 1: The picture is clicked by me.
Where:
$r = 6$
$c = 6$
The n-gram modified precision score of order i, as depicted in equation (4) is computed using Table 1.

Table 1: A summary results for the Computation of n-gram Modified Precision Score

n = 1				n = 2				n = 3				n = 4			
	MT l = 6	Ref r = 6	Min (MT , Ref)		MT	Ref	Min (MT , Ref)		MT	Ref	Min (MT, Ref)		MT	Ref	Min (MT , Ref)
"the"	2	1	1	"the pictur e"	2	1	1	"the picture the"	1	0	0	"the picture the picture"	1	0	0
"picture "	2	1	1	"pictu re the"	1	0	0	"pictur e the picture "	1	0	0	"picture the picture by"	1	0	0
"by"	1	1	1	"pictu re by"	1	0	0	"the picture by"	1	0	0	"the picture by me"	1	0	0
"me"	1	1	1	"by me"	1	1	1	"pictur e by me"	1	0	0				
$p_1 = \frac{4}{6} = \frac{2}{3}$				$p_2 = \frac{2}{5}$				$p_3 = \frac{0}{4} = 0$				$p_4 = \frac{0}{3} = 0$			

Hence, using equation (3), the Brevity Penalty is computed as:

$$BP = \left(1,\ \frac{r}{c}\right) = \left(1,\ \frac{6}{6}\right) = (1,\ 1) = 1$$

Substituting BP with other parameters into equation (1) gives:

$$BLEU\ score = BP * \exp exp \sum_{i=1}^{N} w_i p_i$$

$$BLEU\ score = 1 * \exp exp \sum_{i=1}^{4} w_i p_i$$

$$BLEU\ score = 1 * \exp exp\ (0.25\frac{2}{3} + 0.25\frac{2}{5} + 0 + 00)\ [w_1 = w_2 = 0.25,\ w_3 = 0,\ w_4 = 0];$$

$$BLEU\ score = 0.718 = 71.8$$

2.2.1.2 Character-level BLEU (charBLEU)

This is a BLEU score metric that computes the BLEU score on the character level rather than on the sentence level [6]. It is a better evaluation metric than the ASR BLEU score.

2.2.1.3 Weaknesses of BLEU Score

1. It is an n-gram precision-based metric that does not take into consideration the recall, and its reliance on the exact matching of the n-gram [7].
2. It does not show correlation when compared with human judgment for speech-to-speech translation tasks [7].

2.2.2 Word Error Rate (WER)

For speech-to-speech translation tasks, the WER, an objective evaluation for comparing the ground truth word string to machine translated word string is obtained using equation (5) [16-17] as:

$$WER = \frac{S+D+I}{N} \tag{5}$$

where:

$S = The\ number\ of\ substitutions$
$D = The\ number\ of\ deletions$
$I = The\ number\ of\ insertions$
$N = The\ number\ of\ words\ in\ the\ reference$

2.2.2.1 Computation of WER in Speech-Speech Translation

Given the reference text (ground truth) of the target language (Yorùbá) as GT and the Translated text equivalent of the translated speech of the target language (Yorùbá) of the output of the translator as MT as given below, the WER is computed using equation (13) as illustrated below:

GT: *A bẹ̀rẹ̀ ìmúlò àjẹsára ibà pọ́njú − pọ́ntọ̀ ní ọdún* 1938.

MT: *A bẹ̀rẹ̀ ìmúlò àwọ́n àjẹsá ibà pọ́njú − pọ́ntọ̀ ọdúni .*

The WER is computed using equation (13) as:

$S = 2$ [*àjẹsá for àjẹsára and ọdúni for ọdún*
$D = 2$ [*ní and* 1938 *deleted in the MT obtained from ASR (input from target speech output)*]
$I = 1$ [*àwọ́n inserted in the MT output*]
$N = 9$

$$WER = \frac{S+D+I}{N} = \frac{2+2+1}{9} = \frac{5}{9} = 0.55$$

2.2.3 Recall-Oriented Understudy Gisting Evaluation (ROUGE – L)

This is an evaluation metric that compares the machine translated text sequence (ASR output obtained from speech audio output) with that of the ground truth text sequence by finding the Longest Common Subsequence (LCS) of words. It is mostly used in texts summarisation models like the GPT-4 [18] as well as machine translation [7]. According to findings it is much more efficient to compute the non-contiguous LCS of words than its contiguous counterpart to capture more flexible matches between the ground truth and machine translation texts as the order of words may be different. Statistical metrics such as precision, recall and f1 score can be computed using the LCS of both the ground truth word strings and machine translation word strings.

Given that:

$X = Reference\ or\ ground\ truth\ word\ or\ sequence$

$Y = Machine\ Translation\ word\ or\ sequence\ output$

$LCS(X, Y) = LongestCommon\ Subsequence\ of\ X, and\ Y\ (non\ contiguous)$

The statistical metrics are computed as giving in equations (6) to (8) follows [9]:

$$precision,\ p = \frac{len(LCS(X,Y))}{len(Y)} \tag{6}$$

$$recall,\ r = \frac{len(LCS(X,Y))}{len(X)} \tag{7}$$

$$f1 - score = \frac{\left(\left(1+\beta^2\right) * p * r\right)}{\left(r+\left(\beta^2 * p\right)\right)} \tag{8}$$

where:

$len(LCS(X, Y)) = Length\ of\ the\ LCS\ of\ X, and\ Y$

$len(X) = Length\ of\ the\ ground\ truth\ word\ string\ or\ sequence$

$len(Y) = Length\ of\ the\ machine\ translated\ word\ string\ or\ sequence$

$\beta = Parameter\ that\ is\ chosen\ to\ compute\ the\ f1 - score\ and\ controls\ the\ trade - off\ between$

Note that setting $\beta = 1$ makes equation (8) equals equation (9) (for computation of the statistical, $f1$) as illustrated below:

$$f1 - Score = \frac{2*Precision*Recall}{Precision+Recall}$$

$$f1 - score = \frac{\left(\left(1+1^2\right)*p*r\right)}{\left(r+\left(1^2*p\right)\right)} \tag{9}$$

$$f1 - score = \frac{(2*p*r)}{(r+p)}$$

When the value of $\beta = 1$, the weight of recall, r is equal to that of the precision p. The value β can also be set below or above 1. When $\beta = \frac{1}{2}$, more weights are allocated to the precision, p and this applicable where precision, p is crucial and when $\beta = 2$, more weights are allocated to the recall and it is utilised where recall, r is crucial.

Note that β is also obtained using $\frac{p}{r}$ [9].

2.2.3.1 Computation of Statistical Metrics and F-Based ROUGE-L in Speech-Speech Translation

Given the reference text (ground truth) of the target language (Yorùbá) as X and the Translated text equivalent of the translated speech of the target language (Yorùbá) of the output of the translator as Y as given below, the precision, recall, and F1-Score are calculated for ROUGE-L using equations (6-8) as illustrated below:

X: *A bẹ̀rẹ̀ ìmúlò àjẹsára ibà pọ́njú − pọ́ntọ̀ ní ọdún* 1938.

Y: *A bẹ̀rẹ̀ ìmúlò àjẹsá ibà pọ́njú − pọ́ntọ̀ ọdúni* 1938.

$LCS(X, Y) =$ *A bẹ̀rẹ̀ìmúlò àjẹsá ibà pọ́njú − pọ́ntọ̀ ọdún* 1938. *(non contiguous subsequence)*

$len(X) = 9$

$len(Y) = 8$
$len(LCS(X, Y)) = 8$

$$precision,\ p = \frac{len(LCS(X,Y))}{len(Y)} = \frac{8}{8} = 1$$

$$recall,\ r = \frac{len(LCS(X,Y))}{len(X)} = \frac{8}{9} = 0.88$$

$$f1 - based\ ROUGE - L = \frac{\left((1+\beta^2) * p * r\right)}{\left(r+\left(\beta^2 * p\right)\right)} = \frac{\left((1+1^2) * 1 * 0.88\right)}{\left(0.88+\left(1^2 * 1\right)\right)} = \frac{2*1*0.88}{0.88+1} = \frac{1.77}{1.88} = 0.94$$

Note that in the first example given in this section using equations (6-8), the computation of precision, recall, and f1-score were carried out using word level counting and the parameter, $\beta = 1$

Considering the character level ROUGE – L computation, the following is obtained:

$len(X) = 50$
$len(Y) = 46$
$len(LCS(X, Y)) = 45$

Note that all characters such as alphanumerical, and special characters as well as white spaces are counted as characters.

$$precision,\ p = \frac{len(LCS(X,Y))}{len(Y)} = \frac{45}{46} = 0.97$$

$$recall,\ r = \frac{len(LCS(X,Y))}{len(X)} = \frac{45}{50} = 0.90$$

$$f1 - based\ ROUGE - L = \frac{\left((1+\beta^2) * p * r\right)}{\left(r+\left(\beta^2 * p\right)\right)} = \frac{\left((1+1^2) * 0.97 * 0.90\right)}{\left(0.90+\left(1^2 * 0.97\right)\right)} = \frac{2*0.97*0.90}{0.90+0.97} = \frac{1.746}{1.870} = 0.93$$

2.2.3.2 Advantages of using ROUGE-L for Speech-to-Speech Translation Model

The ROUGE-L score takes into consideration the longest common subsequence between the machine translated output and reference texts. This subsequence can either be contiguous or non-contiguous. For machine translation, utilising the contiguous nature of ROUGE-L ensures it avoids the consecutive matching of words for word level metric or consecutive matching of character for character level matching. This ensures it generalises across the whole texts capturing differences between the machine generated and ground truth texts. For speech-to-speech translation models, it could be used in integration with BLEU, and METEOR to enhance the translation model performance since it is possible for two different texts (sharing relationship to the ground truth text) to have same ROUGE-L score when compared to the ground truth texts [9]. In addition to that, the ROUGE-L score can also be utilised for assessing the quality of the translated texts which is essential in speech-to-speech translation tasks.

2.2.3.3 Limitation of using ROUGE-L for Speech-to-Speech Translation Model

ROUGE-L being a text-based metric, shows it has the capability to introduce errors typical of text-based speech processing metrics that could affect the performance of the model developed. Another limitation to utilising ROUGE-L is its usage in evaluating or comparing two similar machine translated texts obtained from two different models when compared to the ground truth texts. This is because, ROUGE-L cannot tell which one is close to the ground truth, rather it is an analytical approach to computing the performance metric. Such a limitation could be handled by the subjected evaluation using MOSN or MOSS where raters rate the speech output directly. In addition to that, ROUGE-L is not suitable for evaluating the naturalness, or quality of speech as this can be harnessed from the speech

translated output. It is only used to evaluate the quality of the texts which may have deviated from the original speech due to ASR errors.

2.2.4 Mel-Cepstra Distortion (MCD)

This is a performance metric that compares the predicted target mel-cepstra with the reference mel-cepstra [19]. It is calculated as the difference between the MFCCs of the predicted target and reference audio. Mathematically, it is represented in equation (10) as proposed by [20]:

$$MCD_k = \frac{1}{T}\sum_{t=1}^{T}\sqrt{\sum_{k=1}^{K}(m_{tk} - \hat{m}_{tk})^2} \quad (10)$$

where:

m_{tk} = kth MFCC of the $t - th$ frame from the reference audio

$\hat{m}_{tk}$ = kth MFCC of the $t - th$ frame from the predicted audio

Then, MCD_k is the sum of the squared differences over the first K MFCCs. When the length of the two MFCCs sequences are not equal, Dynamic Time Warping (DTW) is utilised to compute the minimum distance MCD obtainable [19-21]. This is used to evaluate speech synthesis models.

2.2.5 BERTScore

This is language generation [39] evaluation metric that is mostly utilized for machine translation. It can also be utilized for speech-to-speech translation models by feeding the output speech obtained to an ASR model and then compares the texts obtained with the reference texts. BERTScore is based on pretrained BERT contextual embeddings. Hence, it computes the cosine similarity of the machine translation texts and ground truth or reference texts-this it does by finding the sum of the cosine similarity between their respective token's embeddings [39]. BERTScore was known to perform better than n-gram metrics such as BLEU score, METEOR, as well as ROUGE-L in machine translation [7]. It addresses two major problems associated with n-gram metrics, which are penalizing semantical-ordering of words, inability to capture distant dependencies, and their inability to match paraphrases. It was confirmed to have evaluation performance close to human judgement [39].
Given:

The reference / ground truth parameters as:

Tokenized sentence: $y = (y_1, y_2, y_3, y_4, \ldots, y_k)$
Embedding vectors of y: $Y = (Y_1, Y_2, Y_3, Y_4, \ldots, Y_k)$

The machine translated output parameters as:

Tokenized sentence: $\hat{y} = (\hat{y}_1, \hat{y}_2, \hat{y}_3, \hat{y}_4, \ldots, \hat{y}_m)$
Embedding vectors of y: $\hat{Y} = (\hat{Y}_1, \hat{Y}_2, \hat{Y}_3, \hat{Y}_4, \ldots, \hat{Y}_m)$,

The recall, precision, and F1-score are computed for BERTScore using equations (11-13) as [39]:

$$R_{BERT} = \frac{1}{|y|} \sum_{y_i \epsilon y} Y_i^T \hat{Y}_j \quad (11)$$

$$P_{BERT} = \frac{1}{|\hat{y}|} \sum_{\hat{y}_{j \epsilon \hat{y}}} Y_i^T \hat{Y}_j \quad (12)$$

$$F_{BERT} = \frac{2 * P_{BERT} * R_{BERT}}{P_{BERT} + R_{BERT}} \quad (13)$$

The computed BERTScore of equations (11-13) are within the cosine similarity range of -1 to 1, which does not affect human correlation or ranking of BERTScore. To ensure human readability, the range is adjusted to fall within 0 and 1. This is carried using empirical lower bound, b that is computed using Common Crawl monolingual datasets. The rescaled BERTScores for equations (11-13) are computed using equations (14-16) as [39]:

$$\hat{R}_{BERT} = \frac{R_{BERT} - b}{1 - b} \quad (14)$$

$$\hat{P}_{BERT} = \frac{P_{BERT} - b}{1 - b} \quad (15)$$

$$\hat{F}_{BERT} = \frac{F_{BERT} - b}{1 - b} \quad (16)$$

It is to be noted that scaling carried out above is an optional step. Prior to this step, another optional step that involves weighting with Inverse Document Frequency (IDF) score can also be done to obtain the weighted BERTScores instead of equation (11-13).

2.2.5.1 Pseudo Code for the Computation of BERTScore for Speech-to-Speech Translation Model

A. Pseudo code for BERTScore (Precision BERTScore)

1. *Begin*
2. Compute the tokenised sequences y, and $\hat{y}$ of both the reference texts and machine translation respectively.
3. Compute the embeddings Y and $\hat{Y}$ of both the reference texts and machine translation respectively.
4. For the first word $\hat{y}_1$ in the machine translation, compute the cosine similarity between its embedding, $\hat{Y}_1$ and all embeddings $Y_1, Y_2, Y_3, Y_4, \ldots, Y_k$ in the reference texts.
5. Compute the maximum value given as: $Y_i^T \hat{Y}_j$
6. Repeat step 5 for the remaining words $\hat{y}_2, \hat{y}_3, \hat{y}_4, \ldots, \hat{y}_m$ in the machine translated output.
7. Compute the sum of the results of step 5 and 6 as: $\sum_{\hat{y}_{j \epsilon \hat{y}}} Y_i^T \hat{Y}_j$
8. Divide the result of step 7 by the total number of tokens in the machine translation output (candidate output) given by $\frac{1}{|\hat{y}|} \sum_{\hat{y}_{j \epsilon \hat{y}}} Y_i^T \hat{Y}_j$ to obtain P_{BERT}.

9. Compute the optional scaling precision score using: $\hat{P}_{BERT} = \frac{P_{BERT} - b}{1 - b}$
10. *End*

B. Pseudo code for BERTScore (Recall BERTScore)

1. *Begin*
2. Compute the tokenised sequences y, and $\hat{y}$ of both the reference texts and machine translation respectively.
3. Compute the embeddings Y and $\hat{Y}$ of both the reference texts and machine translation respectively.
4. For the first word y_1 in the reference texts, compute the cosine similarity between its embedding, Y_1 and all embeddings $\hat{Y}_1$, $\hat{Y}_2$, $\hat{Y}_3$, $\hat{Y}_4$, ..., $\hat{Y}_m$ in the machine translated output.
5. Compute the maximum value given as: $Y_i^T \hat{Y}_j$
6. Repeat step 5 for the remaining words y_2, y_3, y_4, ..., y_k in the reference texts.
7. Compute the sum of the results of step 5 and 6 as: $\sum_{y_i \epsilon y} Y_i^T \hat{Y}_j$
8. Divide the result of step 7 by the total number of tokens in the reference texts given by $\frac{1}{|y|} \sum_{y_i \epsilon y} Y_i^T \hat{Y}_j$ to obtain R_{BERT}.
9. Compute the optional scaling recall score using: $\hat{R}_{BERT} = \frac{R_{BERT} - b}{1 - b}$
10. *End*

2.2.6 Real Time Factor (RTF)

This metric is utilised to calculate the performances of ASR models. It measures the processing speed of audio. A higher RTF value shows a faster processing of audio signals. Mathematically, it is computed as the ratio of processing time to audio duration given in equation (17) as [61, 62]:

$$REAL - TIME\ FACTOR\ (RTF) = \frac{Processing\ Time}{Audio\ Duration} \quad (17)$$

2.2.7 Sort-Time Objective Intelligibility (STOI)

This metric measures the intelligibility of the speech signals. It can be used to evaluate how clean a speech signal is from its degraded replica. It is mostly utilised for text-to-speech models. For speech-to-speech translation task, it can be utilised to find how intelligible the target speech is from its reference speech. It is measured on a scale of 0 to 1 where 0 denotes unintelligibility while 1 means perfect intelligibility. The steps to compute the STOI of a speech signals are as follows [60]:

2.2.7.1 Steps to Compute the STOI of Speech Signal

1. Split the speech signals into short-time frames which is typically between 32 – 64 ns (256 – 512 samples with 50 % overlaps.
2. Calculate the Sort-Time Fourier Transform (STFT) for both the clean and degraded speech signals of each frame.

3. Compute the spectral magnitude of the STFT for each frame of both the clean and degraded speech signals.
4. Normalize the calculated spectral magnitude to have the same energy in each frame for both the clean and degraded speech signals.
5. Compute the correlation coefficients between the clean and degraded speech signals for each frame.
6. Calculate the STOI score by finding the average of step 5 across all frames.

2.2.8 Perceptual Evaluation of Speech Quality (PESQ)

This measures the quality of speech signals obtained from speech models. It is mostly used for speech synthesis models. It can also be used for speech-to-speech translation models. It measures speech quality on a scale of -0.5 to 4.5 where a higher score means better speech quality. A PESQ of -0.5 means bad speech quality while 4.5 PESQ denotes excellent speech [59, 63].

2.2.8.1 Steps to Compute the PESQ

1. Pre-process the clean and degraded speech signals (reference and target speech in the case of speech models) using pre-processing techniques such as filtering, and normalization.
2. Carry out time alignment of both the degraded and clean speech to check for any distortions or delays.
3. The disturbance or degradation which is the difference between the clean and target speech signals is computed.
4. Mapping of the disturbance to a PESQ score that ranges from -0.5 to 4.5 to calculate the score.

2.3 Statistical Evaluation

This refers to the utilisation of machine learning evaluation metrics. This involves Precision, Recall, F1-Score, etc.

2.3.1 Accuracy

This is the ratio of the sum of True Positive (TP) and True Negative (TN) to the sum of TP, TN, False Positive (FP), and False Negative (FN). It is obtained using equation (18) as [23]:

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{18}$$

2.3.2 Precision

This is the ratio of TP to the sum of TP and FP. Mathematically, it is obtained as given in equation (19) as [22-24]:

$$Precision = \frac{TP}{TP+FP} \tag{19}$$

Relating it to speech-to-speech translation tasks, it is computed using equation (20) as [25]:

$$Precision = \frac{No\ of\ correctly\ translated\ words\ (bold)}{Total\ number\ translated\ words} \tag{20}$$

2.3.3 Recall

This is the ratio of TP to the sum of TP and False Negative (FN). Mathematically, it is obtained as given in equation (21) [22-24]:

$$Recall = \frac{TP}{TP+FN} \tag{21}$$

For speech-to-speech translation tasks, it is computed using equation (22) as [25]:

$$Recall = \frac{No\ of\ correctly\ translated\ words\ (bold)}{Total\ number\ reference\ words} \tag{22}$$

2.3.4 F1-Score

This is the ratio of the product of precision and recall to the sum of precision and recall. Mathematically, it is obtained using equation (23) [22-24], which is same as equation (9) as:

$$F1 - Score = \frac{2*Precision*Recall}{Precision+Recall} \tag{23}$$

For speech-to-speech translation tasks, it is also computed using equation (9) which is same as equation (23).

2.3.5 Computation of Precision, Recall, and F1-Score in Speech-Speech Translation

Given the reference text (ground truth) of the target language (Yorùbá) as RT and the Translated text equivalent of the translated speech of the target language (Yorùbá) of the output of the translator as MT as given below, the precision, recall, and F1-Score are calculated using equations (19), (21), and (22) respectively as illustrated below [25].

RT/GT: A bẹ̀rẹ̀ ìmúlò àjẹsára ibà pọ́njú-pọ́ntọ̀ ní ọdún 1938.

MT: A bẹ̀rẹ̀ ìmúlò àjẹsá ibà pọ́njú-pọ́ntọ̀ ọdúni 1938.

$$No\ of\ correctly\ translated\ words\ (bold) = 6$$

$$Total\ number\ translated\ words\ in\ the\ output\ of\ the\ translator = 8$$

$$Total\ number\ of\ reference\ words = 9$$

$$Precision = \frac{No\ of\ correctly\ translated\ words\ (bold)}{Total\ number\ translated\ words} = \frac{6}{8} = 0.75$$

$$Recall = \frac{No\ of\ correctly\ translated\ words\ (bold)}{Total\ number\ reference\ words} = \frac{6}{9} = 0.66$$

$$F1 - Score = \frac{2*Precision*Recall}{Precision+Recall} = \frac{2*0.75*0.66}{0.75+0.66} = 0.702$$

III. COMPARATIVE ANALYSIS OF THE DIFFERENT EVALUATION METRICS

Table 2 shows the comparison of some of the speech-to-speech translation metrics so that researchers can make right choices for their speech-to-speech models. It details the subjective and objective evaluation metrics, such as MOS Naturalness, MOS Similarity, BLASER, & XSTS; and BLEU score, WER, METEOR, ROUGE-L, BERTScore, & MCD respectively. The table also highlights the statistical

metric that could also be explored for speech-to-speech translation tasks [7, 9]. Table 3 gives highlights of when, application areas and how each of the discussed metrics are utilised.

Table 2: Comparison of Various Performance Metrics for Speech-to-Speech Translation Tasks

Evaluation Metrics	Application	Fluency	Adequacy	Audio translation quality	Naturalness
Subjective Evaluations					
MOS Naturalness	Target output speech			Excellent	Excellent
MOS Similarity	Target output speech	Excellent	Excellent		
BLASER	Target output speech			Excellent. It is text-free. [12]	
XSTS	Target output speech		Excellent	Excellent [12]	
Objective Evaluations					
BLEU score	Text obtained from Target output speech	Excellent [9]	Excellent [9]	Excellent [12]	
ROUGE-L	Text obtained from Target output speech	Excellent	Good		
WER	Text obtained from Target output speech	Excellent [9]	Good [9]		
MCD	Cepstral features of both source and target speech			Excellent	
Statistical Metrics (precision, recall, and f1-score)	Text obtained from Target output speech	Excellent	Good		
METEORS	Text obtained from Target output speech	Excellent [7]		It has high correlation with human judgment.	It has high correlation with human judgment.
BERTScore	Text obtained from Target output speech	Excellent	Excellent		

Table 3: When, where, and how to use the Speech Translation Metrics

Performance metrics	When	Where	How
MOS Naturalness	1. Fluency, adequacy, and audio quality are needed.	Speech-to-speech translation, automatic speech translation, speech synthesis tasks	Sourcing human raters to judge the target speech using a scale of 1 to 5 where 1 = Poor, 2 = Satisfactory, 3

			= Good, 4 = Very good, and 5 = Excellent.
MOS Similarity	1. Fluency, adequacy, and audio quality are needed.	Speech-to-speech translation, automatic speech translation, speech synthesis tasks	Same as above
BLASER	1. Audio quality is to be tested.	Speech-to-speech translation tasks	Translated output speech is converted into vectors and the vectors are then fed into a small dense neural network for prediction of XSTS scores for each output of the translation for the supervised version of BLASER 2.0
XSTS	1. Adequacy, and audio quality are needed.	Speech-to-speech translation tasks.	It is used by human raters who are bilingual and judge how adequate the target speech is to be source speech on using a score on a scale of 1 to 5.
BLEU score	1. Fluency, adequacy, and audio quality are needed.	Speech-to-speech translation, automatic speech translation, speech synthesis, machine translation tasks	Computed on n-gram words, and evaluated using equations (1) to (4)
ROUGE-L	1. Non-contiguous subsequence order of words or character is needed. 2. Quality of texts is needed.	Speech-to-speech translation, machine translation, texts summarisation	Computed on word-to-word matching. It is evaluated using equations (6) to (9)
WER	1. Quality of texts is needed.	Speech-to-text, text-to-speech, speech-to-speech translation tasks.	Computed on texts using equation (5)
MCD	1. Audio quality is needed.	Speech synthesis tasks	Computed on texts using equation (10)
Statistical Metrics (precision, recall, and f1-score)	1. Quality of texts is needed.	Speech-to-speech translation tasks,	Computed on texts using equation (17) to (22)
METEORS	1. When word-to-word matching is needed between the reference and machine translated output. 2. Word order is needed.	Speech-to-speech translation tasks	Computed on the texts obtained from the ASR output fed with target speech.
BERTScore	1. Word ordering and capturing of dependencies is needed. 2. A good score for paraphrasing is needed.	Speech-to-speech translation tasks, sentence summarizations,	Computed on the contextualized embeddings between the reference and machine translated texts. It

		machine translation tasks.	is evaluated using equations (11) to (16)
RTF	1. Processing speed of target audio signal is needed	Speech-to-text	Computed using equation (17)
STOI	1. Speech intelligibility is needed.	Text-to-speech, speech-to-speech translation tasks	Computed using the steps highlighted in section 2.2.7
PESQ	1. Speech quality is needed	Text-to-speech, speech-to-speech translation tasks	Computed using the steps highlighted in section 2.2.8

IV. STATE OF THE ART (SOTA) SPEECH-TO-SPEECH MODELS' PERFORMANCE METRICS

The performance metrics utilised by some SOTA speech-to-speech translation models metrics utilised by researchers in this field are highlighted in Table 4. It shows that some researchers utilised the BLEU score along with MOS for speech-to-speech translation tasks. Some new metrics like the BLASER, XSTS, and ROUGE – L are beginning to be used by researchers in this area of research including speech-to-texts, and automatic speech-to-texts translation [27].

Table 4: Speech-to-Speech Translation Metrics Utilised for well-known SOTA models

Performance Metrics / Speech Models / Authors	BLEU scores	MOS Naturalness	MOS Similarities	Statistical Precision, Recall, f1-score	ROUGE-L / MCD / WER / BLASER	Others
Ref. [11]		MOS	MOS		WER	
Ref. [19]	BLEU score	Yes				
Ref. [20]	BLUE	MOS	MOS		WER	Latency
Ref. [28] - AudioPalm	ASR BLEU	Yes	Yes		WER	
Ref. [4] – Translatotron	BLEU score	Yes	Yes			Phoneme Error Rate (PER)
Ref. [29] - Translatotron2	BLEU score	Yes	Yes			
Ref [26] - Whisper	BLEU score		Yes			
Ref. [21]	BLEU score	MOS	MOS			
Ref. [27]	BLEU-4, Google BLEU			Yes	ROUGE-L	METEOR. BERTScore
Ref. [12, 30]	ASR BLEU	MOS Naturalness		ASR chrF		BLASER 2.0, XSTS, percentage acceptable translation, METEOR
Ref. [15]	BLEU score	Yes				

Ref. [31]	BLEU score					
Ref. [32]	BLEU score					Character Error Rate (CER)
Ref. [33]	ASR BLEU					
Ref. [34]	Units-BLEU, ASR BLEU	MOS	SMOS	ASR chrF	BLASER 2.0	Speaker Encoder Cosine Similarity (SECS)
Ref. [35]	BLEU					Character Error Rate (CER)
Ref. [36]	ASR BLEU					
Ref. [16]					WER, ROUGE-L	
Ref. [17]					WER	Word Recognition Rate (WRR), RTF
Ref. [8]	BLEU					
Ref. [37]					MCD	

V. APPLICATIONS OF SPEECH METRICS ON LEADERBOARDS AND MODERN AI APPLICATIONS

In this modern-day era, speech models are mostly compared with other models to know how they perform among themselves. Different models are evaluated or tested using the same dataset and ranked based on their performance. For instance, speech-to-speech translation models such as Whispers [26], Translatotron, SeamlessM4T may be benchmarked with same datasets of different varieties, across different databases such as FLUERS, LibriSpeech, MustC, Fisher datasets, and then the BLEU scores are computed for each model across these datasets and are ranked based on these metrics. When more than one source of datasets is utilized, the average across these sources is computed to rank each model.

At present, there are quite a few leaderboards and modern AI platforms [45-46] that evaluate and rank speech models such as the Hugging Face, International Workshop on Spoken Language Translation (IWSLT), Real-World Speech-to-Text API Leaderboard, Open ASR leaderboard etc [40-41, 42-43, 46-48]. Metrics such as BLEU scores, WER, METEORS, Real Time Factor (RTF), Short-Time Objective Intelligibility (STOI), Perceptual Evaluation of Speech Quality (PESQ), CER, are being utilized to evaluate and rank speech models. There is no known information about whether subjective metrics such as MOS Naturalness, MOS Similarity, BLASER, and MCD are being used on any of these platforms. BERTScore, METEOR, ROUGE-L, and BLEU scores are used on Hugging Face for texts generation models like texts summarization, and machine translation models. On the Open ASR Leaderboard on Hugging Face, ASR models are being compared and ranked, and metrics of choice are the WER, and Real Time Factor (RTF). Where the WER and RTF are utilized to evaluate the performance, the speech-to-texts models and are ranked based on these metrics. However, this review paper is suggesting the utilization of BERTScore for speech-to-speech translation models where it can

be used on the transcribed texts of the target speech. Then the text is compared to the reference text by computing the cosine similarity between them. In such application, it is important to not rely on the BERTScore alone but to integrate it with other metrics such as the subjective evaluation metrics, and the BLEU score because the BERTScore cannot evaluate speech quality even though it evaluates the word order. Some metrics are peculiar to each leaderboard as well as speech models. For instance, on the Hugging Face leaderboard, A detailed overview of these leaderboards is given in the next sub section.

5.1 Hugging Face Leaderboard

Hugging Face leaderboard is an AI platform that evaluates the SOTA AI models such as the image processing models, text-based models, and speech processing models [42-43, 46-48, 50-51]. The platform evaluates and compares ASR models on its Open ASR Leaderboard [50]. ASR models are ranked using the WER, and RTF. TTS models are also being ranked via this platform [47-48, 51].

5.2 ICASSP 2024 Speech Signal Improvement Challenge

This is a speech models competing leaderboard that aims to improve speech signals by evaluating speech models for their speech signals quality and intelligibility. Speech signals are evaluated using techniques such as filtering, noise reduction, and speech enhancement. The metrics utilized here are Short Time Objective Intelligibility (STOI), and Perceptual Evaluation of Speech Quality (PESQ), Word Accuracy (WAcc) [65].

5.3 IberSPEECH 2024 Challenge

This platform was developed to promote research and development in speech processing applications such as speech synthesis, ASR etc. It looks into the evaluation of speech recognition and speech synthesis models. WER, and Character Error Rate (CER) are used to evaluate the ASR models while text-to-speech models are evaluated using PESQ, and STOI [64].

5.4 International Workshop on Spoken Language Translation (IWSLT)

This platform promotes research in Spoken Language Translation (SLT) such as speech-to-text, speech-to-speech translation, automatic speech translation, Speech synthesis etc., [40-41]. Speech-to-texts and speech-to-speech translation models are evaluated and ranked on this platform where BLEU scores, and METEOR are utilized to evaluate the ASR models, while speech-to-speech translation models are evaluated using BLEU scores, METEOR, and WER.

5.5 Speech Generation Evaluation and Leaderboard

This platform evaluates and ranks speech generation models using metrics such as speech intelligibility, which is measured by speech recognition error rates; Naturalness, which is predicted utilising speech models trained on human naturalness ratings; and Similarity, which measures the cosine similarity speaker embeddings mostly used for voice cloning systems [46, 49].

VI. TYPICAL LEADERBOARD RATINGS FOR DIRECT AND CASCADED SPEECH-TO-SPEECH TRANSLATION MODELS.

6.1 ASR BLEU Speech-to-Speech Translation on FLEURS X-ENG

For the ASR BLEU metric, the speech-to-speech translation tasks from other language to English on FLUERS corpus shows that GenTranslateV2, GenTranslateV1, SeamlessM4T LargeV2, SeamlessM4T Large, AudioPaLM2, WhisperV2, SeamlessM4T Medium have ASR BLEU scores of 32.3, 30.1, 29.4,

25.8, 24.0, 23.5, and 20.4 respectively, where the highest BLEU score of 32.3, was obtained for GenTranslateV2, with SeamlessM4T Medium having the lowest BLEU score. This leaderboard is available on this link: https://paperswithcode.com/sota/speech-to-speech-translation-on-fleurs-x-eng?utm_source=chatgpt.com. It should be noted that the authors of the GenTranslateV2 developed both end-to-end and cascaded system where the end-to-end models performed better than all the other models over 30 languages to English translation on both FLEURS X-Eng and CoVoST X-Eng datasets. For 15 languages to English on FLEURS X-Eng dataset, it achieved an average BLEU score of 32.3 while GenTranslateV1, SeamlessM4T LargeV2, SeamlessM4T Large, AudioPaLM2, and WhisperV2 have average BLEU scores of 30.1, 29.4, 27.1, 24.0, 23.5 respectively. For the cascaded system, GenTranslateV2, GenTranslateV1, SeamlessM4T V2, SeamlessM4T, and WhisperV2 have average BLEU scores of 34.2, 34.0, 32.3, 31.9, and 31.2 respectively. This shows GenTranslateV2 still performs better in the cascaded speech-to-speech translation task [52].

6.2 ASR BLEU For Speech-to-Speech Translation on CVSS dataset

On another rank, where both SeamlessM4T Large and SeamlessM4T Medium were ranked, results on the leaderboard shows that SeamlessM4T Large had the best ASR BLEU with a value of 36.5 in comparison with 28.1 for SeamlessM4T Medium when both were trained on CVSS Dataset. The link is available here: https://paperswithcode.com/sota/speech-to-speech-translation-on-cvsss

6.3 ASR WER Speech-to-Text Translation on Hugging Face Leaderboard

Using the 8 datasets used in the ESB paper [53], which consists of LibriSpeech clean, LibriSpeech other, VoxPopuli, TED-LIUM, GigaSpeech, SPGISpeech, Earnings-22, and AMI datasets as the benchmark datasets, Granite-speech-3.3-8b which was trained on public and synthetically generated datasets for ASR, and Automatic Speech Translation (AST) tasks achieved the best WER of 5.85 in comparison to Massively Multilingual Speech (MMS) - Finetuned ASR - FL102 with a wave2vec architecture which has the worst WER of 39.8. SOTA Whisper large, and Whisper medium have WERs of 7.94, and 8.09 respectively. Whisper-large-v3 has the best WER amongst all the Whisper models ranked with a value of 7.44. Facebook's Hubert-xlarge-ls960-ft has WER of 22.55. It should be noted that all models ranked on this leaderboard were trained with the same 8 training datasets aforementioned after which comparison were made. The average WER across all the 8 datasets were computed for each ASR model [54].

6.4 WER for ASR on TedLium Dataset

In another leaderboard for ASR WER on TedLium dataset (a dataset that contains English language TED Talks which spans from 118 to 452 hours sampled at 16 kHz with their respective transcripts), United-MedASR trained on 764 parameters, parakeet-rnnt-1.1b, Whispering-LLaMa-7b, and SpeechStew trained 100 M parameters yielded WERs of 0.29, 3.92, 4.6, and 5.3 respectively, which shows that United-MedASR has the best WER. This leaderboard is available at this link: https://paperswithcode.com/sota/speech-recognition-on-tedlium

VII. CASCADED & DIRECT SPEECH-TO-SPEECH PERFORMANCE METRICS

The conventional approach to speech-to-speech translation models before the exploration of machine / deep learning-based methods involves utilising cascading of traditional statistical approaches such as Hidden Markov Model (HMM), Forward Algorithms, Viterbi Algorithm etc. for ASR, statistical approaches such as HMM, analysis of transcribed words for machine translation and concatenative approach, rule-based approach, statistical approach for text-to-speech [4, 20, 55-56]. At a later time when AI is evolving, the machine learning based approach involving deep learning such as neural

networks, LSTM etc., [20, 57-58] are being applied in this cascaded approach to modeling speech-to-speech translation task. Due to the availability of high computing power [61, 62], coupled with big data, deep learning based a direct end-to-end approach begins to take over the training of speech-to-speech translation model, where texts representation as seen in the cascaded approach is not available [4]. Of course, one, would expect that metrics of evaluation for the cascaded approach will involve integration of metrics for ASR, MT, and SS models, whereas the metrics of the direct approach involves a single metrics per time to evaluate target speech output. Table 5 highlights some of the metrics utilised for cascaded and direct speech-to-speech translation tasks.

Table 5. Summary of Cascaded and Direct Speech-to-Speech Translation Metrics

Performance Metrics Speech Models / Authors	Approach	Speech-to-Speech Metrics Used
Ref [1]	Direct	BLEU, METEOR
Ref. [19]	Cascaded (MT, speech synthesis, speech recognition)	MT- BLEU score, ChrF, CharBLEU, Speech synthesis - MOS Naturalness, MOS Similarity, MCD
Ref. [20]	Cascaded	Streaming ASR (WER), Simultaneous MT (BLEU score), Incremental Multi-lingual TTS (MOS Naturalness, MOS Similarity), Latency
Ref. [28] – AudioPalm	Direct	ASR BLEU, MOS Similarity, BLEU, WER
Ref. [4] – Translatotron	Direct, (compared with cascaded)	BLEU score, MOS Naturalness, Phoneme Error Rate (PER)
Ref. [29] - Translatotron2	Direct	BLEU score, MOS Naturalness, MOS Similarity
Ref [26] – Whisper	Direct / Self-supervised	BLEU score, MOS Similarity
Ref. [27]	Cascaded and Direct	BLEU-4, Google BLEU, MOS Similarity, ROUGE-L, METEOR. BERTScore, NIST
Ref. [12, 30] - Seamless M4T	Direct	ASR BLEU, MOS Naturalness), ASR chrF, BLASER 2.0, XSTS, percentage acceptable translation, METEOR
Ref. [15]	Cascaded	BLEU score, MOS Naturalness, MOS Similarity,
Ref. [31]	Direct (compared with cascaded)	BLEU score
Ref. [32]	Direct	BLEU score, Character Error Rate (CER)
Ref. [33]	Direct	ASR BLEU
Ref. [34]	Direct	Units-BLEU, ASR BLEU, MOS, SMOS, Speaker Encoder Cosine Similarity (SECS)
Ref [10]	Direct	ASR BLEU
Ref [6]	Cascaded (compared with Direct -Translatotron)	BLEU Scores, MOS Naturalness

VIII. DISCUSSION

The study highlights the various performance metrics utilised by researchers for speech processing applications, particularly speech-to-speech translation tasks. Based on the findings as reported in this work, most of the existing speech-to-speech translation models have utilised objective BLEU score performance metric for evaluation [7, 9, 15, 30-33]. MOS Naturalness, and MOS Similarity follow next as the performance metrics in this domain [12, 28-29, 34]. Other metrics that could be utilised for

speech-to-speech translation tasks are METEOR, and ROUGE-L [1], [27]. Other metrics such as the WER, CER, NIST, XSTS, SER, PER, and BLASER are also of interest in this field of speech processing applications.

A thorough analysis of the various performance metrics utilised for speech-to-speech translation tasks show that all the subjective evaluations are computed on the generated output speech while that of the objective evaluation metrics show they are computed on the transcribed text string using a STT model except the MCD which is mostly utilised for TTS [35-37]. Hence, it is expected that the subjected evaluation metrics gives the best performance when compared to its objective counterpart due to no errors introduced as a result of the absence of ASR transformation [19]. Findings show that when it comes to adequacy which shows how close the predicted target output speech is to the source speech, MOS Similarity as well as XSTS, BLEU score, and ROUGE-L could be utilised to perform evaluation. MOS Naturalness can be utilised to check the naturalness of the output speech. For audio translation quality, the MOS Naturalness, BLASER, XSTS, as well as BLEU score could be utilised to achieve that [12]. The outcome of the study further suggests that ROUGE-L performance metric and its utilization in the NLP tasks such as machine translation and speech recognition [7], [27] should be given priority as an evaluation metric for speech-to-speech translation tasks. This is due to the fact that it is to be applied on the transcribed texts [1] obtained from the generated output target speech just like the case of the BLEU score metric. However, care must be taken due to its limitation of not being able to give good evaluations of two or more different translated texts that have same number of words (word-level) or characters (character-level) but different order - same ROUGE-L scores [9]. It is therefore advisable that it is integrated with other metrics such as BLEU, and METEOR. Coupled with that, ASR chrF is known to perform better than ASR BLEU due to its ability to perform more matching between the translated output texts and ground truth texts at the character level. WERs are mostly utilised for ASR models, but they can also be used for speech-to-speech translation tasks. MCD has little application for speech translation tasks. They are mostly utilised for speech synthesis from texts (TTS model) [36-38].

The study went further to survey some of the present leaderboards and modern AI tools available for evaluating and comparing speech models [44-45]. Some of the existing leaderboards such as Facebook's Hugging Face, IWSLT etc., [40-41, 42-43] evaluates speech models such as the TTS, and ASR. These evaluations give experts insight into how the speech models perform in comparison to one another. There are also leaderboards specifically for TTS and speech generations [49, 50] that compare TTS models with one another. The choice of metrics utilized in any of the leaderboards is dependent on the speech models being considered, as well as the leaderboards of interest. For instance, on Facebook's Open ASR leaderboards, the WER, and RTF are being used to compare and evaluate ASR models.

IX. CONCLUSION

The various metrics utilised for speech-to-speech translation models, their benefits and the comparison of the metrics were highlighted in this study. The results showed that among the various speech models metrics that have been employed, BLEU score is predominant, followed by MOS Naturalness and MOS Similarity. Metrics such as the BLASER, XSTS, ROUGE-L, BERTScores, and METEORS are beginning to gain recognition particularly for speech models. Leaderboards ratings of speech models show that existing BLEU scores, WER, and RTF are mostly utilized for speech models. This further suggests the importance of these metrics. RTF even though it is being utilized on the leaderboards, has the least utilization by researchers. Also, metrics such as ROUGE-L, METEORS, and BERTScore have been utilised successfully in machine translation, they could be explored for speech-to-speech translation particularly when they are used in conjunction with each other coupled with BLEU score. The findings show that the subjective metrics are computed directly on the target speech output, which depicts a good evaluation of the quality of speech model, unlike the objective metrics that are prone to errors

introduced due to the ASR utilization. The study therefore serves as an eye opener to researchers or readers hoping to develop speech-to-speech translation models to make the right choices on the performance metrics of their models.

ACKNOWLEDGEMENTS

The authors wish to acknowledge the Covenant University Center for Research, Innovation and Discovery (CUCRID), and Google for providing funds towards the publication of this study

REFERENCES

1. Tjandra Andros, Sakriani Sakti, & Satoshi Nakamura (2019) "Speech-to-speech translation between untranscribed unknown languages," in *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 593–600.
2. Salesky Elizabeth, Julian Mäder, & Severin Klinger (2021) "Assessing evaluation metrics for speech-to-speech translation," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 733–740.
3. Zhang Chen, Xu Tan, Yi Ren, Tao Qin, Kejun Zhang, & Tie-Yan Liu (2021) "Uwspeech: Speech to speech translation for unwritten languages," in *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 14319–14327.
4. Jia Ye, Ron J. Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, & Yonghui Wu (2019) "Direct speech-to-speech translation with a sequence-to-sequence model," In *Proc. Interspeech,* pp. 1123-1127.
5. Kano Takatomo, Sakriani Sakti, & Satoshi Nakamura (2021) "Transformer-based direct speech-to-speech translation with transcoder," in *2021 IEEE spoken language technology workshop (SLT)*, pp. 958–965.
6. Dong Qianqian, Fengpeng Yue, Tom Ko, Mingxuan Wang, Qibing Bai, & Yu Zhang (2022) "Leveraging pseudo-labeled data to improve direct speech-to-speech translation," In *Proc. Interspeech*, pp. 1781-1785. 2022
7. Blagec Kathrin, Georg Dorffner, Milad Moradi, Simon Ott, & Matthias Samwald (2022) "A global analysis of metrics used for measuring performance in natural language processing," In *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*, pp. 52-63.
8. Kala, R. Jules, Emmanuel Adetiba, Abdultaofeek Abayom, Oluwatobi E. Dare, & Ayodele H. Ifijeh (2025) "Speech to Speech Translation with Translatotron: A State of the Art Review," *arXiv preprint arXiv:2502.05980*
9. Lin Chin-Yew, & Franz Josef Och (2004) "Orange: a method for evaluating automatic evaluation metrics for machine translation," in *COLING 2004: Proceedings of the 20th International Conference on Computational Linguistics*, pp. 501–507.
10. Inaguma Hirofumi, Sravya Popuri, Ilia Kulikov, Peng-Jen Chen, Changhan Wang, Yu-An Chung, Yun Tang, Ann Lee, Shinji Watanabe, & Juan Pino. (2023) "Unity: Two-pass direct speech-to-speech translation with discrete units," In *The 61st Annual Meeting of The Association For Computational Linguistics*.
11. Licht Daniel, Cynthia Gao, Janice Lam, Francisco Guzman, Mona Diab, & Philipp Koehn (2022) "Consistent human evaluation of machine translation across language pairs," In *Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pp. 309-321.
12. Barrault Loïc, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar et al. (2023) "SeamlessM4T-Massively Multilingual & Multimodal Machine Translation," *arXiv preprint arXiv:2308.11596*.

13. Dale David, & Marta Costa-jussà (2024) “Blaser 2.0: a metric for evaluation and quality estimation of massively multilingual speech and text translation,” in *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 16075–16085.
14. Chen Mingda, Paul-Ambroise Duquenne, Pierre Andrews, Justine Kao, Alexandre Mourachko, Holger Schwenk, & Marta R. Costa-jussà (2023) “BLASER: A Text-Free Speech-to-Speech Translation Evaluation Metric,” In *The 61st Annual Meeting Of The Association For Computational Linguistics*.
15. Gong Zairan, Xiaona Xu, & Yue Zhao (2025) “Tibetan–Chinese speech-to-speech translation based on discrete units,” *Sci Rep*, Vol. 15, No. 1, pp. 2592, 2025.
16. Medeiros Eduardo Farófia (2023) “Deep learning for speech to text transcription for the Portuguese language.”
17. Malik Mishaim, Muhammad Kamran Malik, Khawar Mehmood, & Imran Makhdoom (2021) “Automatic speech recognition: a survey,” *Multimed Tools Appl*, Vol. 80, No. 6, pp. 9411–9457, doi: 10.1007/s11042-020-10073-7.
18. Nelson Max, Shannon Wotherspoon, Francis Keith, William Hartmann, & Matthew Snover (2024) “Cross-Lingual Conversational Speech Summarization with Large Language Models,” *arXiv preprint arXiv:2408.06484*.
19. Salesky Elizabeth, Julian Mäder, & Severin Klinger (2021) “Assessing Evaluation Metrics for Speech-to-Speech Translation,” [Online]. Available: http://arxiv.org/abs/2110.13877
20. de Martos, Alejandro Manuel Pérez González (2022) “Deep neural networks for automatic speech-to-speech translation of open educational resources,” Universitat Politècnica de València.
21. Müller Meinard (2007) “Dynamic time warping,” *Information retrieval for music and motion*, pp. 69–84, 2007.
22. Subramanian Barathi, Rathinaraja Jeyaraj, Rakhmonov Akhrorjon Akhmadjon Ugli, & Jeonghong Kim (2024) “Enhancing Sequential Model Performance with Squared Sigmoid TanH (SST) Activation Under Data Constraints.”
23. Ndolu Fajar Henri Erasmus, & Ruki Harwahyu (2023) “Intrusion Detection System on Nowadays’ Attack using Ensemble Learning,” *IJNMT (International Journal of New Media Technology)*, Vol. 10, No. 1, pp. 42–50.
24. Okokpujie Kennedy, Daniella Kalimumbalo, Joke A. Badejo, & Emmanuel Adetiba (2022) “Congestion Intrusion Detection-based Method for Controller Area Network Bus: A case for KIA SOUL Vehicle. "*Mathematical Modelling of Engineering Problems* 9, Vol. 5.
25. Hujon, V. Aiusha, Thoudam Doren Singh, & Khwairakpam Amitab (2022) “Transfer Learning Based Neural Machine Translation of English-Khasi on Low-Resource Settings,” in *Procedia Computer Science*, Elsevier B.V., 2022, pp. 1–8. doi: 10.1016/j.procs.2022.12.396.
26. Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, Ilya Sutskever, “Robust speech recognition via large-scale weak supervision,” in *International Conference on Machine Learning*, 2023, pp. 28492–28518.
27. Khurana Sameer, Nauman Dawalatabad, Antoine Laurent, Luis Vicente, Pablo Gimeno, Victoria Mingote, & James Glass. (2023) “Improved cross-lingual transfer learning for automatic speech translation,” *arXiv preprint arXiv:2306.00789*.
28. Rubenstein K. Paul, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen et al. "Audiopalm: A large language model that can speak and listen." *arXiv preprint arXiv:2306.12925* (2023). (2023) “AudioPaLM: A Large Language Model That Can Speak and Listen,”, [Online]. Available: http://arxiv.org/abs/2306.12925
29. Jia Ye, Michelle Tadmor Ramanovich, Tal Remez, & Roi Pomerantz (2022) “Translatotron 2: High-quality direct speech-to-speech translation with voice preservation.” In *International Conference on Machine Learning*, pp. 10120-10134.

30. "Joint speech and text machine translation for up to 100 languages," *Nature*, Vol. 637, No. 8046, pp. 587–593, 2025.
31. Nguyen Luan Thanh, & Sakriani Sakti (2025) "ZeST: A Zero-resourced Speech-to-Speech Translation Approach for Unknown, Unpaired, and Untranscribed Languages," *IEEE Access*, 2025.
32. Shi Jiatong, Yun Tang, Ann Lee, Hirofumi Inaguma, Changhan Wang, Juan Pino, & Shinji Watanabe *(2023)*, "Enhancing speech-to-speech translation with multiple TTS targets," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5.
33. Diwan Anuj, Anirudh Srinivasan, David Harwath, & Eunsol Choi (2023) "Textless low-resource speech-to-speech translation with unit language models," *arXiv e-prints*, p. arXiv–2305.
34. Song Kun, Yi Ren, Yi Lei, Chunfeng Wang, Kun Wei, Lei Xie, Xiang Yin, & Zejun Ma (2023) "StyleS2ST: Zero-shot Style Transfer for Direct Speech-to-speech Translation," [Online]. Available: http://arxiv.org/abs/2305.17732
35. Ogayo Perez, Graham Neubig, & Alan W. Black (2022) "Building African Voices," Jul. 2022, [Online]. Available: http://arxiv.org/abs/2207.00688
36. Ogunremi Tolulope, Kola Tubosun, Anuoluwapo Aremu, Iroro Orife, & David Ifeoluwa Adelani (2023) "Iroyin Speech: A multi-purpose Yoruba Speech Corpus,", [Online]. Available: http://arxiv.org/abs/2307.16071
37. Gutkin Alexander, Isin Demirsahin, Oddur Kjartansson, Clara Rivera, & Kólá Túbòsún. (2020) "Developing an Open-Source Corpus of Yoruba Speech", In *Interspeech,* pp. 404-408
38. Meyer Josh, David Ifeoluwa Adelani, Edresson Casanova, Alp Öktem, Daniel Whitenack Julian Weber, Salomon Kabongo, Elizabeth Salesky et al. (2022) "BibleTTS: a large, high-fidelity, multilingual, and uniquely African speech corpus", In *Proc. Interspeech*, pp. 2383-2387
39. Zhang Tianyi, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. (2019) "BERTScore: Evaluating Text Generation with BERT." In *International Conference on Learning Representations*.
40. Mathias Müller *et al.* (2022) "Findings of the first WMT shared task on sign language translation (WMT-SLT22)," in *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pp. 744–77
41. Deepanjali Singh, Ayush Anand, Abhyuday Chaturvedi, and Niyati Baliyan (2024) "IWSLT 2024 Indic Track system description paper: Speech-to-Text Translation from English to multiple Low-Resource Indian Languages," in *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pp. 311–316.
42. U. R. Pol, P. S. Vadar, and T. T. Moharekar, "Hugging Face: Revolutionizing AI and NLP".
43. Y. Shen, K. Song, X. Tan, D. Li, W. Lu, and Y. Zhuang, "Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face," *Adv Neural Inf Process Syst*, vol. 36, pp. 38154–38180, 2023.
44. Data Science Dojo. (2025). *Top 5 LLM Leaderboard Platforms for AI Excellence*. [online] Available at: https://datasciencedojo.com/blog/understanding-llm- leaderboards/ [Accessed 23 Jun. 2025].
45. Artificialanalysis.ai. (2025). *Speech to Speech Models and Providers Analysis | Artificial Analysis*. [online] Available at: https://artificialanalysis.ai/models/speech-to-speech.
46. Balacoon (2025). *Speech Generation Evaluation and Leaderboard*. [online] Balacoon. Available at: https://balacoon.com/blog/tts_leaderboard/ [Accessed 23 Jun. 2025].
47. AGI, T. (2025). *TTS Arena Legacy*. [online] Huggingface.co. Available at: https://huggingface.co/spaces/TTS-AGI/TTS-Arena [Accessed 23 Jun. 2025].
48. Huggingface.co. (2025). *TTS Spaces Arena - a Hugging Face Space by Pendrokar*. [online] Available at: https://huggingface.co/spaces/Pendrokar/TTS-Spaces-Arena.
49. *Speech generation | Leaderboard*. (2025). Labelbox.com. https://labelbox.com/leaderboards/speech-generation/?utm_source=chatgpt.com

50. Face, H. (2025). *Open ASR Leaderboard*. [online] Huggingface.co. Available at: https://huggingface.co/spaces/hf-audio/open_asr_leaderboard?utm_source=chatgpt.com [Accessed 24 Jun. 2025].
51. AGI, T. (2025). *TTS Arena V2*. [online] Huggingface.co. Available at: https://huggingface.co/spaces/TTS-AGI/TTS-Arena-V2 [Accessed 24 Jun. 2025].
52. Y. Hu *et al.*, "GenTranslate: Large language models are generative multilingual speech and machine translators," *arXiv preprint arXiv:2402.06894*, 2024.
53. S. Gandhi, P. Von Platen, and A. M. Rush, "Esb: A benchmark for multi-domain end-to-end speech recognition," *arXiv preprint arXiv:2210.13352*, 2022.
54. Face, H. (2025). *Open ASR Leaderboard*. [online] Huggingface.co. Available at: https://huggingface.co/spaces/hf-audio/open_asr_leaderboard?utm_source=chatgpt.com.
55. Nallabala, N. K., Souprayen, B., Ramasamy, M., Penumarti, S. K., & Navuluri, N. C. (2025). An Efficient Speech Synthesizer: A Hybrid Monotonic Architecture for Text-to-speech via VAE & LPC-Net with Independent Sentence Length. *Circuits, Systems, and Signal Processing*, 1–25.
56. Medeiros, E. F. (2023). *Deep learning for speech to text transcription for the portuguese language*. Universidade de Évora.
57. L. Santos, N. de Araújo Moreira, R. Sampaio, R. Lima, and F. C. M. B. Oliveira, "Automatic Speech Recognition: Comparisons Between Convolutional Neural Networks, Hidden Markov Model and Hybrid Architecture," *Expert Syst*, vol. 42, no. 5, p. e70032, 2025.
58. S. Sultoni, B. Darmawan, and others, "Speaker Recognition System Using MFCC and HMM Methods," *International Journal of Informatics and Computation*, vol. 7, no. 1, pp. 206–218, 2025.
59. tsbmail (2024). *P.862:Perceptual evaluation of speech quality (PESQ): An objective method for end-to-end speech quality assessment of narrow-band telephone networks and speech codecs*. [online] Itu.int. Available at: https://www.itu.int/rec/T-REC-P.862/en [Accessed 24 Jun. 2025].
60. Taal, C.H., Hendriks, R.C., Heusdens, R. and Jensen, J. (2010). *A short-time objective intelligibility measure for time-frequency weighted noisy speech*. [online] IEEE Xplore. doi:https://doi.org/10.1109/ICASSP.2010.5495701.
61. N. Zheng, X. Wan, K. Liu, and Z. Huan, "SCDiar: a streaming diarization system based on speaker change detection and speech recognition," *arXiv preprint arXiv:2501.16641*, 2025
62. M. Tran *et al.*, "A Domain Adaptation Framework for Speech Recognition Systems with Only Synthetic data," *arXiv preprint arXiv:2501.12501*, 2025.
63. M. Torcoli, M. M. Halimeh, and E. A. P. Habets, "Navigating PESQ: Up-to-Date Versions and Open Implementations," *arXiv preprint arXiv:2505.19760*, 2025.
64. Isca-archive.org. (2024). *ISCA Archive*. [online] Available at: https://www.isca-archive.org/iberspeech_2024/index.html [Accessed 25 Jun. 2025].
65. Cutler, Ross, Ando Saabas, Tanel Pärnamaa, Marju Purin, Evgenii Indenbom, Nicolae-Cătălin Ristea, Jegor Gužvin, Hannes Gamper, Sebastian Braun, and Robert Aichner. (2024) "ICASSP 2023 acoustic echo cancellation challenge." *IEEE Open Journal of Signal Processing*.