Re-evaluating Minimum Bayes Risk Decoding for Automated Speech Recognition Tasks

Anonymous authors
Paper under double-blind review

Abstract

While sample-based Minimum Bayes Risk (MBR) decoding has shown to outperform beam search in many text-to-text generation tasks with modern LLMs, beam search remains the dominant approach for Automatic Speech Recognition (ASR) and Speech Translation (ST). To date, the efficacy of MBR decoding within modern speech systems lacks comprehensive evaluation. Given that MBR decoding is effective in text-to-text generation tasks, it is reasonable to expect it to also be effective for speech-to-text tasks. In this paper, we evaluate MBR decoding for ASR and ST tasks on English and Japanese using Whisper and its derivative models. We observe that the accuracy of MBR decoding outperforms that of beam search in most of the experimental settings we have evaluated. The results show that MBR decoding is a promising method for ASR and ST tasks that require high accuracy. The anonymized code is available at `https://anonymous.4open.science/r/mbr-for-asr-56F0`.

1 Introduction

Automatic Speech Recognition (ASR) is the task of converting spoken language into written text and plays a crucial role in a wide range of applications. Advances in deep learning have significantly improved the accuracy and robustness of ASR systems, enabling their deployment in diverse real-world scenarios (Prabhavalkar et al., 2024).

Decoding algorithms play an important role in determining the final output quality of ASR systems. One of the common approaches, beam search, incrementally explores the most probable hypotheses to approximate the maximum-a-posteriori (MAP) solution. While effective and efficient, beam search is known to suffer from several degeneration issues in text-to-text generation tasks such as machine translation (Holtzman et al., 2020; Eikema & Aziz, 2020). Minimum Bayes risk (MBR) decoding offers a promising alternative by directly optimizing for the expected utility of the output (Goel & Byrne, 2000; Kumar & Byrne, 2004). Rather than selecting the single most probable sequence, MBR considers multiple candidate hypotheses and chooses the one that minimizes the expected loss (or maximizes utility) when compared against other likely outputs (Bickel & Doksum, 2015). This approach has shown remarkable success in text-to-text tasks such as machine translation, summarization, and captioning (Eikema & Aziz, 2022; Suzgun et al., 2023; Jinnai et al., 2024; Wu et al., 2025), consistently outperforming beam search across diverse evaluation metrics.

While MBR decoding has been evaluated for classic ASR systems (e.g., hidden Markov model, Goel & Byrne 2000; Goel et al. 2004), its application to speech-to-text tasks with the modern ASR systems has not been investigated (Prabhavalkar et al., 2024). For example, MBR decoding has been applied to the spoken language translation in the recent IWSLT shared tasks (Ahmad et al., 2024; Abdulmumin et al., 2025), but it is used for the machine translation modules rather than the ASR modules of the cascaded systems (Yan et al., 2024; Ben Kheder et al., 2024; Li et al., 2024; Wang et al., 2025; Romney Robinson et al., 2025).

Given that the method is designed to improve the decoding accuracy of probabilistic models in general (Ichihara et al., 2025a), it is reasonable to expect it to also improve the accuracy of ASR modules. The absence of comprehensive studies on MBR decoding for contemporary ASR systems represents a significant

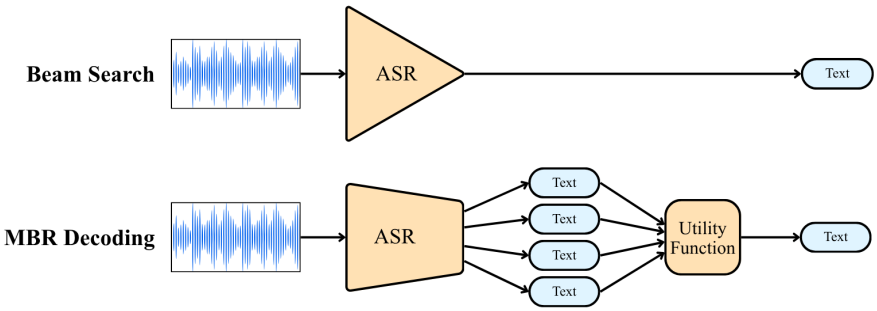

Figure 1: Illustration of the beam search and MBR decoding: multiple hypotheses are sampled from the ASR model, and the hypothesis with the highest expected utility (e.g., BLEU score) compared to the others is selected as the final output.

gap in the literature. Given MBR's empirical successes in text-to-text tasks and theoretical advantages, a systematic evaluation of its potential for speech recognition is valuable.

To this end, we present a comprehensive evaluation of sample-based MBR decoding for both offline ASR and Speech Translation (ST) tasks (Figure 1). Our experiments span multiple languages, with a focus on English and Japanese. We use diverse datasets, multiple models based on Whisper, and with varying levels of synthesized noise added. MBR decoding consistently outperforms beam search across these dimensions, often by substantial margins. Remarkably, these improvements emerge with as few as 4-8 samples, suggesting that MBR can be practically implemented in scenarios where latency requirements are not stringent.

Our findings have significant implications for high-accuracy ASR applications where transcription quality takes precedence over real-time processing. While the computational overhead of MBR makes it less suitable for real-time applications, its consistent accuracy improvements make it an attractive option for offline speech-to-text systems. This work thus reestablishes MBR decoding as a valuable technique in the modern neural ASR toolkit.

2 Background

We first formally define the text generation problem and then describe MBR decoding.

2.1 Text Generation Problem

Conditional text generation is the problem of generating a sequence of tokens $y \in \mathcal{Y}$ conditioned on an input context x, using a probabilistic model $P(y|x)$, where $\mathcal{Y}$ is the set of all possible sequences (Graves, 2012; Sutskever et al., 2014). Formally, we denote $\mathcal{V}$ as a set of tokens (vocabulary). Let `bos` and `eos` be special tokens representing the beginning and end of a sequence, respectively. Then, $\mathcal{Y}$ is the set of sequences of tokens from the vocabulary $\mathcal{V}$, starting with `bos` and ending with `eos`:

$$\mathcal{Y} = \{(\texttt{bos}, y_1, y_2, \ldots, y_n, \texttt{eos}) | n \geq 0, y_i \in \mathcal{V}\}. \tag{1}$$

The context x can be any modality, such as text (i.e., $x \in \mathcal{Y}$), image, and audio. The tasks include important real-world problems such as machine translation, image captioning, and ASR, where the goal is to produce an output sequence that is appropriate given the input.

A straightforward solution is a maximum a posteriori (MAP) estimate, which selects the most likely output sequence given the input context:

$$\hat{y}_{\text{MAP}} = \arg\max_{y \in \mathcal{Y}} P(y|x). \tag{2}$$

Given that $\mathcal{Y}$ is typically very large in text generation tasks, it is often infeasible to enumerate all possible output sequences in $\mathcal{Y}$. Thus, local optimal search methods such as beam search are used to approximate the MAP estimate as the language models are typically modeled by a autoregressive models (Vaswani et al.,

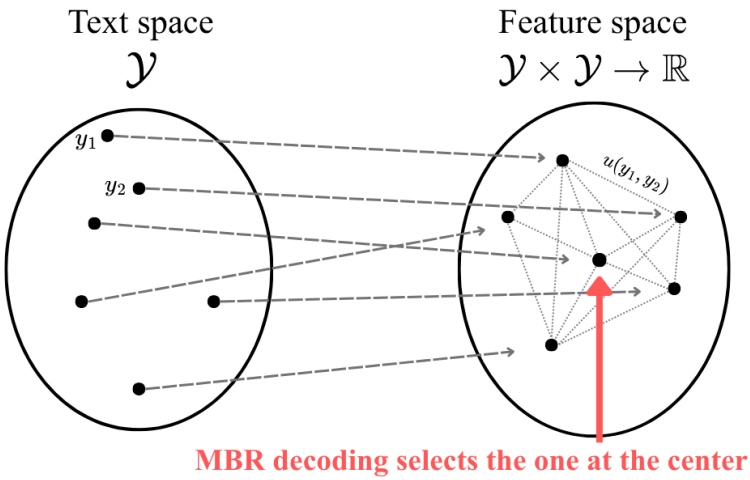

Figure 2: Illustrative explanation of the intuition of the MBR decoding. The hypothesis that lies at the center of the sampled hypotheses is selected as the output. The distance between two hypotheses is inversely related to their utility.

2017). However, MAP decoding, including beam search, is known to generate undesirable outputs, such as an empty sequence, a sequence with repeated tokens, or low-quality text (Wiseman et al., 2017; Holtzman et al., 2020; Eikema & Aziz, 2020). Thus, alternative decoding algorithms have been investigated to improve the quality of the generated text.

2.2 Minimum Bayes Risk (MBR) Decoding

MBR decoding works by sampling multiple hypotheses from the model and selecting the one that maximizes the expected utility compared to the rest of the hypotheses (Goel & Byrne, 2000; Kumar & Byrne, 2004; Eikema & Aziz, 2022):

$$\hat{y} = \arg\max_{y \in H} \frac{1}{N} \sum_{y' \in H} u(y, y'), \tag{3}$$

where H is the set of hypotheses sampled from the model, $N = |H|$ is the number of hypotheses, and $u(y, y')$ is a utility function that measures the quality of hypothesis y by treating y' as a pseudo-reference drawn from the set of peer hypotheses. Intuitively, MBR decoding selects the hypothesis that lies at the *center* of the sampled hypotheses, where the *distance* between two hypotheses is inversely related to their utility (Figure 2).[1] It should be noted, however, that this "center" reflects the model's own distribution: if the model generates systematically biased outputs (e.g., a tendency toward shorter sentences), MBR will select a hypothesis near the center of that biased distribution and thus will not correct for such model-intrinsic biases. Whereas MAP decoding selects the sequence with the highest probability in the discrete hypothesis space, MBR decoding selects the one that lies near the middle of the continuous space defined by the utility function. The utility function implicitly defines a continuous space over the hypotheses by quantifying their pairwise similarities.

Previous studies have shown that the way hypotheses H are sampled is crucial for MBR decoding performance (Eikema & Aziz, 2022; Suzgun et al., 2023; Jinnai et al., 2024; Ohashi et al., 2024). Originally, Goel & Byrne (2000) proposed MBR decoding for ASR using beam search to generate H, and this was later applied to machine translation (Kumar & Byrne, 2004). However, recent work has found that using unbiased samples drawn from the model is more effective than using beam search for generating H (Eikema & Aziz, 2022). Other studies have also reported that probabilistic sampling methods, such as ancestral sampling, nucleus

[1]The utility functions used in MBR decoding are often not symmetric and may not satisfy the triangle inequality, so they are not proper distance functions. Nevertheless, the intuition still holds in many practical cases.

sampling (Holtzman et al., 2020), and epsilon sampling (Hewitt et al., 2022; Freitag et al., 2023), work better than beam search (Eikema & Aziz, 2022; Ohashi et al., 2024).

Another advantage of MBR decoding is that it has theoretical guarantees (Ichihara et al., 2025a). Under mild assumptions, the expected utility of the output chosen by MBR decoding improves as the number of sampled hypotheses increases, with a rate of $O(\frac{1}{\sqrt{N}})$. This result is consistent with empirical findings showing that larger sample sizes lead to higher generation quality (Freitag et al., 2023). In contrast, beam search lacks non-vacuous theoretical guarantees regarding output quality.

The drawback of MBR decoding is its computational cost. The complexity is $O(UN^2 + GN)$, where U is the cost of computing the utility function and G is the cost of generating a hypothesis (Eikema & Aziz, 2022). There are faster algorithms (Cheng & Vlachos, 2023; Deguchi et al., 2024; Trabelsi et al., 2024) that reduce the cost to $O(UN \log N + GN)$ (Jinnai & Ariu, 2024), but this is still much higher than beam search, which is $O(GB)$, where B is the beam width. Note that G represents the computational cost of a full decoding step per hypothesis including any pruning operations as constant factors.

In summary, MBR decoding is a strong alternative to beam search for text generation tasks, and it consistently performs better in many settings. It is not only effective in practice but also has theoretical support. Its main weakness is its computational cost, which makes it less suitable for real-time use. There has been little evaluation of MBR decoding for speech-to-text tasks, which this paper aims to address.

3 Related Work

MBR decoding can be understood as a generalized form of *consensus decoding* as it selects the hypothesis that minimizes expected risk with respect to the distribution of sampled hypotheses, effectively finding the consensus of the hypotheses. When the utility function is a token-level edit metric (such as WER or Levenshtein distance), MBR reduces to a form of ROVER-style majority voting (Fiscus, 1997), which directly optimizes for word error rate rather than maximizing the posterior probability (i.e., MAP decoding). Consensus decoding, in turn, can be understood as an instance of reranking methods (Morbini et al., 2012; Chiu & Chen, 2021; Xu et al., 2022; Nakano et al., 2022; Ichihara et al., 2025b) which rank the hypotheses according to its utility and select the best one. Several reranking methods have been proposed for speech recognition using quality estimation (Negri et al., 2014; Ng et al., 2015; Ali & Renals, 2018; Yuksel et al., 2023; Waheed et al., 2025), perplexity (Salazar et al., 2020), deliberation models (Hu et al., 2020; Xu et al., 2022), LLMs (Nie et al., 2022; Hu et al., 2024; Tur et al., 2024), and speech-text foundational models (Shivakumar et al., 2025). The advantage of MBR decoding compared to these approaches is that it does not require any additional training, making it easy to apply to new systems and languages.

Model fusion is another approach to improve the accuracy of ASR and ST systems by combining multiple models (Parikh et al., 2024). This approach has been shown to be effective in various settings, such as combining acoustic models and language models (Lei et al., 2023; Chen et al., 2024) and combining multiple ASR systems (Fiscus, 1997; Tan et al., 2020; Kamo et al., 2025). MBR decoding can be seen as a form of model fusion. In fact, several studies have proposed using MBR decoding to ensemble the outputs from multiple systems (Xu et al., 2010; 2011). At the same time, model fusion can be seen as complementary to MBR decoding, as it focuses on improving the underlying model rather than the decoding process.

Post-editing and error correction are alternative approaches that have been proposed to further improve the accuracy of ASR and speech translation outputs (Liu et al., 2020; Kamiya et al., 2021; Leng et al., 2021; Yang et al., 2023; Ma et al., 2023; Radhakrishnan et al., 2023; Chen et al., 2023). These approaches use language models to correct errors in the initial hypotheses, generating a new hypothesis using the language model (Guo et al., 2019; Hrinchuk et al., 2020; Radhakrishnan et al., 2023). This approach is orthogonal to MBR decoding, as it focuses on refining the output after generation rather than re-evaluating multiple hypotheses during decoding.

Model	whisper-small			whisper-medium			whisper-large-v3		
Metric	WER↓	MetricX↓	SemDist↓	WER↓	MetricX↓	SemDist↓	WER↓	MetricX↓	SemDist↓
Beam ($B = 1$)	0.067	2.091	0.085	0.087	1.818	0.073	0.036	1.744	0.064
Beam ($B = 5$)	0.075	2.084	0.085	0.078	1.832	0.075	0.037	1.731	0.064
Beam ($B = 20$)	0.085	2.069	0.086	0.059	1.833	0.075	0.036	1.743	0.062
MBR ($N = 4$)	0.054	1.923	0.074	0.044	1.693	0.066	0.031	1.660	0.058
MBR ($N = 8$)	0.051	1.875	0.070	0.039	1.647	0.061	0.030	1.638	0.056
MBR ($N = 16$)	0.050	1.837	0.068	0.039	1.627	0.061	0.029	1.643	0.055
MBR ($N = 32$)	0.050	1.823	0.068	**0.039**	1.626	0.060	0.029	1.631	0.053
MBR ($N = 64$)	**0.049**	**1.815**	**0.066**	0.040	**1.625**	**0.059**	**0.029**	**1.631**	**0.053**
Model	distil-large-v3.5			s2t-small-librispeech-asr			seamless-m4t-v2-large		
Metric	WER↓	MetricX↓	SemDist↓	WER↓	MetricX↓	SemDist↓	WER↓	MetricX↓	SemDist↓
Beam ($B = 1$)	0.040	1.838	0.056	0.045	2.258	0.042	0.035	1.850	0.031
Beam ($B = 5$)	0.039	1.841	0.055	0.043	2.202	0.039	0.037	1.833	0.029
Beam ($B = 20$)	0.038	1.844	0.057	**0.042**	**2.201**	0.039	0.063	**1.833**	0.029
MBR ($N = 4$)	0.035	1.774	0.049	0.051	2.411	0.046	0.040	1.926	0.033
MBR ($N = 8$)	0.033	1.757	0.047	0.047	2.304	0.042	0.037	1.880	0.030
MBR ($N = 16$)	0.033	1.753	0.045	0.045	2.267	0.040	0.035	1.866	0.029
MBR ($N = 32$)	0.033	1.749	0.045	0.043	2.287	0.040	0.035	1.864	0.029
MBR ($N = 64$)	**0.033**	**1.749**	**0.044**	**0.042**	2.281	**0.040**	**0.034**	1.867	**0.029**

Table 1: Evaluation of beam search and MBR decoding on the full LibriSpeech Clean test set with six models. No noise is synthesized in the audio.

4 Experiments

The goal of the study is to evaluate MBR decoding for ASR and ST tasks, compared to beam search. We investigate various settings, including different models, datasets, and levels of noise added to the input audio.

Method. We conduct experiments to evaluate the performance of MBR decoding and beam search on various ASR and speech translation tasks. For evaluating the methods under noise, we use the free-sound subset of the Musan dataset (Snyder et al., 2015) to induce background noise to the audio. We sample a noise randomly from the freesound subset of the dataset and crop it to match the length of the input audio. The cropped noise audio is synthesized to the speech with the level of Signal-to-Noise Ratio (SNR) set to 0 dB, noted otherwise. The same noise is used for all the decoding algorithms for fair comparison.

For beam search, we run with a beam width of 1, 5, and 20. We generate up to 64 samples for MBR decoding as hypotheses using Epsilon sampling (Hewitt et al., 2022; Freitag et al., 2023) with $\epsilon = 0.01$ and a temperature set to 1.0. We use the BLEU score (Papineni et al., 2002) implemented by the sacrebleu package (Post, 2018) as the utility function of MBR. We do not use WER (CER) as the utility function because MBR decoding is known to inflate the score used as the utility function which may not accurately reflect a model's true capabilities (Freitag et al., 2022; Kovacs et al., 2024). BLEU scores are computed on the normalized texts using Whisper's normalizer for English (Radford et al., 2023) and `neologdn` normalizer for Japanese (Sato et al., 2017) to avoid unnecessary penalization on punctuation. We use MeCab tokenizer (Kudo, 2005) to tokenize Japanese text for computing the BLEU score.

Implementation. All the code of the experiments is implemented by Python 3 using Huggingface's `transformers` library (Wolf et al., 2020). The experiments are conducted on Linux Ubuntu 22.04 using NVIDIA A100 GPUs. While the codebase is not optimized for efficiency, we report the walltime with our implementation as a reference in Section A.

Decoding configuration. For beam search, we use the standard implementation provided in the HuggingFace's `transformers` library. The beam width specifies the number of active hypotheses maintained at each decoding step. No additional pruning strategies are applied: in particular, we do not use threshold

Dataset	LibriSpeech			VoxPopuli			AMI-IHM		
Metric	WER↓	MetricX↓	SemDist↓	WER↓	MetricX↓	SemDist↓	WER↓	MetricX↓	SemDist↓
Beam ($B = 1$)	0.081	2.250	0.091	0.117	1.500	0.067	**0.380**	2.280	**0.264**
Beam ($B = 5$)	0.081	2.230	0.092	0.117	1.500	0.067	**0.380**	2.280	**0.264**
Beam ($B = 20$)	0.082	2.200	0.090	0.117	1.500	0.067	**0.380**	2.280	**0.264**
MBR ($N = 64$)	**0.057**	**2.000**	**0.077**	**0.098**	**1.400**	**0.053**	0.568	**2.200**	0.284

Table 2: Evaluation of beam search and MBR decoding on English ASR tasks with Whisper-large-v3. Noise is synthesized to the audio. The signal-to-noise ratio is 0 dB.

Dataset	ReazonSpeech			CommonVoice			JSUT		
Metric	CER↓	MetricX↓	SemDist↓	CER↓	MetricX↓	SemDist↓	CER↓	MetricX↓	SemDist↓
Beam ($B = 1$)	0.305	2.975	0.143	0.306	2.825	0.134	0.183	2.250	0.088
Beam ($B = 5$)	0.307	2.975	0.143	0.302	2.875	0.132	0.185	2.350	0.089
Beam ($B = 20$)	0.308	3.050	0.140	0.306	2.875	0.133	0.184	2.350	0.090
MBR ($N = 64$)	**0.291**	**2.875**	**0.130**	**0.297**	**2.725**	**0.123**	**0.177**	**2.200**	**0.082**

Table 3: Evaluation of decoding methods on Japanese ASR tasks with Kotoba-Whisper. Noise is synthesized to the audio. The signal-to-noise ratio is 0 dB.

pruning (discarding hypotheses based on a score gap from the best path) or diversity penalties.[2] All reported beam-search results therefore reflect the library defaults beyond the beam width itself.

4.1 Automatic Speech Recognition (ASR)

Resources. We evaluate the performance of MBR decoding on ASR using LibriSpeech (clean) (Panayotov et al., 2015), AMI-IHM (Carletta, 2007), and VoxPopuli (Wang et al., 2021a) for English, ReazonSpeech (Yin et al., 2023), Common Voice-v8 (Ardila et al., 2020), and JSUT (Sonobe et al., 2017) for Japanese. We use Whisper (Radford et al., 2023)[3] for English and Kotoba-Whisper-v2[4] for Japanese ASR models. All the audio files are resampled to 16 kHz to meet the Whisper model's requirement. For the results on Table 1, we use the full test set of LibriSpeech dataset, which contains 2,620 samples, including those longer than 30 seconds. For samples exceeding 30 seconds, because Whisper models do not natively handle audio longer than 30 seceonds, we apply the sequential long-form algorithm provided by the Whisper model (Chiu et al., 2019; Narayanan et al., 2019; Koluguri et al., 2024). For the other experiments, we use the first 1000 samples in the test set for the evaluation, excluding samples longer than 30 seconds to isolate the effect of MBR decoding from any interaction with long-form handling techniques. Most of the samples are shorter than 30 seconds, and the exclusion rate is negligible across these datasets (e.g., only 9 out of 2620 are longer than 30 seconds in LibriSpeech).

Evaluation metrics. We use word error rate (WER) for English and character error rate (CER) for Japanese as the main evaluation metrics. The same normalizers as BLEU scores are used for WER (whisper normalizer) and CER (`neologdn` normalizer). In addition, SemDist (Kim et al., 2021) and MetricX (`metricx-23-xxl-v2p0`; Juraska et al. 2023) are used to evaluate the semantic similarity and overall quality of the generated outputs. SemDist is a metric that measures the semantic distance between the generated text and the reference text using the inner product of the embeddings of the texts, which is also known as other names such as cosine distance and contextual similarity in the NLP community (Akula & Garibay, 2022; Mukherjee & Shrivastava, 2022). It is proposed to complement the problem of WER (CER), which does not capture semantic similarity well, and thus, the effectiveness of the generation in the downstream tasks is not clear by itself. We use a sentence BERT model named `all-MiniLM-L6-v2` as the embedding

[2]Specifically, it follows the implementation and default parameters of `transformers` version 4.55.0.`https://github.com/huggingface/transformers/blob/v4.55.0/src/transformers/models/whisper/generation_whisper.py`

[3]`https://huggingface.co/openai/whisper-large-v3`

[4]`https://huggingface.co/kotoba-tech/kotoba-whisper-v2.0`

SNR (dB)	-20	-15	-10	-5	0	5	10	15	20
Beam ($B = 1$)	0.590	0.458	0.290	0.143	0.081	0.055	0.049	0.049	0.045
Beam ($B = 5$)	0.599	0.446	0.293	0.152	0.081	0.056	0.048	0.049	0.045
Beam ($B = 20$)	0.590	0.444	0.284	0.151	0.082	0.056	0.049	0.048	0.045
MBR ($N = 64$)	**0.530**	**0.388**	**0.235**	**0.108**	**0.057**	**0.041**	**0.035**	**0.036**	**0.034**

Table 4: WER scores on the LibriSpeech dataset with different SNR levels of the speech compared to the synthesized noise.

SNR (dB)	0	5	10	15	20
Beam ($B = 1$)	0.305	0.273	0.258	0.243	0.243
Beam ($B = 5$)	0.307	0.282	0.254	0.243	0.240
Beam ($B = 20$)	0.308	0.272	0.254	0.242	0.235
MBR ($N = 64$)	**0.291**	**0.250**	**0.238**	**0.229**	**0.223**

Table 5: CER scores on the ReazonSpeech dataset with different SNR levels of the speech compared to the synthesized noise.

model to compute SemDist (Reimers & Gurevych, 2019).[5] MetricX is one of the state-of-the-art metrics for machine translation that evaluates the overall quality of the generated outputs by learning human MQM evaluation results. We use it for assessing the overall quality of the generated outputs.

Models. We use whisper-small, whisper-medium, whisper-large-v3, and distil-whisper to evaluate the effect of the model size. Table 1 shows that MBR decoding outperforms beam search in all the model sizes. The result shows that MBR decoding is effective regardless of the model size. We note that beam search results are identical across all tested beam widths ($B = 1, 5, 20$) for the Whisper models evaluated on LibriSpeech (Table 1), VoxPopuli and AMI-IHM (Table 2), and all seven languages in Table 8. We have verified these results against our experimental logs and confirmed their correctness. This behavior stems from the highly peaked probability distributions of the Whisper model family (Radford et al., 2023): the model is extremely confident in its predictions, so the greedy path ($B = 1$) already coincides with the beam-optimal path even for larger beam widths.

We additionally evaluate two non-Whisper sequence-to-sequence models on the LibriSpeech Clean test set: `facebook/s2t-small-librispeech-asr` (S2T; Wang et al. 2020) and `facebook/seamless-m4t-v2-large` (SeamlessM4T; Communication et al. 2023). MBR decoding achieves competitive or better performance than beam search for both models, confirming that the advantage of MBR generalizes across different autoregressive architectures. SeamlessM4T shows larger variation across beam widths (e.g., WER degrades with wider beam), while MBR consistently selects a reliable hypothesis. We also attempted to apply MBR decoding to Wav2Vec 2.0 (Baevski et al., 2020), a CTC-based model. However, the per-frame probability distributions

[5] `https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2`

Metric	WER↓	MetricX↓	SemDist↓
Beam ($B = 1$)	0.042	1.750	0.060
MBR ($N = 64$, $u =$ BLEU)	**0.033**	**1.650**	0.053
MBR ($N = 64$, $u =$ BLEURT)	0.035	**1.650**	0.056
MBR ($N = 64$, $u =$ SentBERT)	0.034	1.675	**0.050***

Table 6: Evaluation of MBR decoding with varying utility functions on the LibriSpeech dataset with whisper-large-v3. No noise is synthesized in the audio. *MBR using SentBERT may lead to inflate SemDist scores that do not accurately reflect a model's true capabilities (Kovacs et al., 2024).

Metric	WER↓	MetricX↓	SemDist↓
Beam ($B = 1$)	0.042	1.750	0.060
MBR ($N = 64$, $\epsilon = 0$)	0.033	**1.625**	0.052
MBR ($N = 64$, $\epsilon = 0.01$)	0.033	1.650	0.053
MBR ($N = 64$, $\epsilon = 0.02$)	**0.033**	1.650	**0.052**

Table 7: Evaluation of MBR decoding with varying sampling parameters on the LibriSpeech dataset with whisper-large-v3. No noise is synthesized in the audio.

Domain	Arabic	Chinese	Hindi	Indonesian	Tamil	Thai	Vietnamese
Beam ($B = 1$)	0.305	0.231	0.180	0.090	0.278	0.366	0.193
Beam ($B = 5$)	0.305	0.231	0.180	0.090	0.278	0.366	0.193
Beam ($B = 20$)	0.305	0.231	0.180	0.090	0.278	0.366	0.193
MBR ($N = 64$)	**0.259**	**0.205**	**0.142**	**0.069**	**0.260**	**0.354**	**0.180**

Table 8: WER (CER) scores of MBR decoding and beam search on the CommonVoice dataset with Whisper-large-v3. No noise is synthesized in the audio. CER is reported for Chinese and WER for other languages.

of CTC models are extremely peaked, so random sampling yields the same result as greedy (MAP) decoding unless the temperature is raised to a level that severely degrades output quality. We therefore conclude that MBR decoding is most effective for models with sufficiently diverse output distributions.

Number of samples for MBR decoding. Table 1 shows the performance of MBR decoding with different numbers of samples. Surprisingly, with only four to eight samples, MBR decoding outperforms beam search. The result shows that MBR decoding is effective even with a small amount of additional computation, which might be admissible for real-time ASR tasks. Still, we observe that the accuracy of MBR decoding improves with a larger number of samples, suggesting that more computation can lead to better performance.

Correlation of MBR objective values to error rates. To investigate how much the MBR objective indicates the utility of the given hypothesis, we compute the correlation of the MBR objective with the WER. Pearson correlation coefficient is computed for each instance of the LibriSpeech with no synthesized noise over 64 samples generated by whisper-large-v3. Then, we estimate it with the average over the 1000 instances. The average value of the Pearson correlation coefficient is -0.3913, and the standard error is 0.0129, indicating that the MBR objective has a substantial negative correlation with the target objective (negative correlation because MBR objective is higher the better, and WER is lower the better). This suggests that it is a reasonable approach to use it as the reranking procedure for ASR.

Datasets. Tables 2 and 3 show the performance of the decoding algorithms using WER (CER), SemDist, and MetricX. MBR decoding outperforms beam search in all the datasets except for AMI-IHM, suggesting that the advantage of MBR decoding over beam search is in a wide range of domains and on both lexical and semantic levels.

Speech length. Given that MBR decoding fails to improve on the AMI-IHM dataset, we conduct a post-hoc error analysis. The corpus records all meeting utterances, resulting in many very short transcriptions that contain non-lexical fillers (e.g., *yeah*, *hmm*, *gosh*). To investigate MBR's performance on these instances, we compute the average WER of beam search and MBR decoding, split by the number of words in the reference transcription (Figure 3). While MBR decoding has a comparable WER to beam search overall, it shows a significantly higher WER on instances shorter than six words. One of the reasons is likely due to the limitations of BLEU on short sequences. If a single filler token is missed or substituted (e.g., *yeah* vs. *yes*), BLEU yields zero overlap, creating a flat utility landscape where no hypothesis is clearly distinguished.

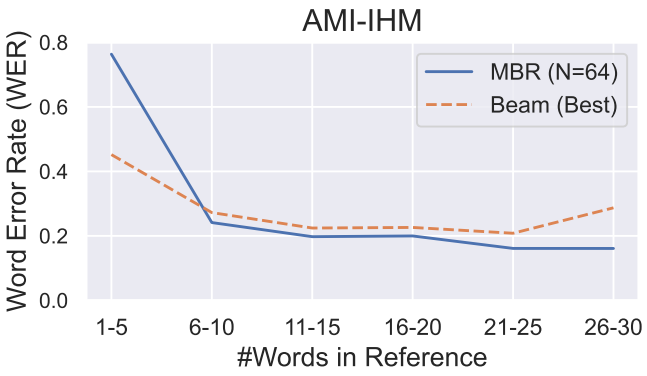

Figure 3: WER of AMI-IHM averaged over the instances with the number of words in the reference text is in the range of (x, x+5].

In this case MBR objective is not informative for selecting the hypothesis, and thus, MBR decoding fails to improve the performance over beam search.

Noise level. Tables 4 and 5 show the performance under different noise levels. The result shows that MBR decoding is more accurate than beam search at any noise level.

Utility functions for MBR decoding. The performance of MBR decoding is known to be dependent on the choice of the utility function (Freitag et al., 2022; Kovacs et al., 2024). We evaluate MBR decoding using SentBERT (Reimers & Gurevych, 2019; 2020) and BLEURT (`BLEURT-20-D12`) in addition to using BLEU. SentBERT is a sentence-level embedding model that captures semantic similarity between sentences computed by the cosine similarity between the two embedding vectors. Thus, the value is 1 minus the value of SemDist. We use the `all-MiniLM-L6-v2` model as the embedding model to compute SentBERT. Table 6 shows that the differences in accuracy using these utility functions are marginal, yet they all outperform beam search. The result shows that the advantage of MBR decoding over beam search is robust to the choice of the utility function. SentBERT achieves the best SemDist score, which is expected as it is directly optimized for the metric (Freitag et al., 2022).

Sampling algorithm for MBR decoding. The choice of sampling algorithm is known to be crucial for the performance of MBR decoding in machine translation tasks (Freitag et al., 2023; Ohashi et al., 2024; Jinnai et al., 2024). We evaluate epsilon sampling with varying epsilon values of 0.00, 0.01, and 0.02. Table 7 shows the performance of MBR decoding with the different epsilon values. The result shows that the performance of MBR decoding is relatively robust to the choice of epsilon values, and it outperforms beam search in all the settings. It also indicates that the effective sampling strategy for ASR may be different from the effective strategy for machine translation (i.e., epsilon sampling), which may be an interesting avenue of future work. We attribute this robustness to the high-confidence nature of current ASR models: in ASR, the model is constrained to transcribe specific acoustic features, resulting in a sharply peaked output distribution where a small number of tokens dominate the probability mass. This makes sampling parameters such as ϵ less consequential than in machine translation, where greater lexical diversity (synonyms, rephrasing) leads to a flatter distribution. This observation is likely tied to the choice of model and task domain, and may not hold universally across all ASR systems.

Languages. To assess whether the performance of MBR decoding is language-specific or generic to natural language text generation tasks, we conduct experiments on the following languages: Arabic (ar), simplified Chinese (zh-CN), Hindi (hi), Tamil (ta), Thai (th), and Vietnamese (vi). We use the test split of the CommonVoice-v8 dataset and evaluate the WER (CER for Chinese). We use spaCy-Thai for segmenting

	LibriSpeech	ReazonSpeech
Beam ($B = 1$)	0.042	0.305
NoRefER ($N = 64$)	0.073	0.368
ProGRes ($N = 64$)	0.043	0.358
MBR ($N = 64$)	**0.033**	**0.291**
Oracle ($N = 64$)	0.013	0.149

Table 9: WER (CER) of the reranking algorithms.

Domain	CoVoST2 (Ja-En)				FLEURS (Ja-En)			
Metric	BLEU↑*	ROUGE-L↑	BLEURT↑	MetricX↓	BLEU↑*	Rouge-L↑	BLEURT↑	MetricX↓
Beam ($B = 1$)	18.646	42.283	-0.184	2.825	6.218	30.178	-0.486	6.750
Beam ($B = 5$)	18.685	41.825	-0.201	2.850	6.158	29.954	-0.487	6.725
Beam ($B = 20$)	18.122	41.362	-0.235	2.950	6.202	29.886	-0.489	6.825
MBR ($N = 64$)	**22.456**	**47.572**	**-0.073**	**2.475**	**8.078**	**34.212**	**-0.365**	**6.100**
Domain	CoVoST2 (En-Ja)				FLEURS (En-Ja)			
Metric	BLEU↑*	Rouge-L↑	BLEURT↑	MetricX↓	BLEU↑*	Rouge-L↑	BLEURT↑	MetricX↓
Beam ($B = 1$)	10.395	34.176	0.110	4.975	8.242	30.771	-0.015	8.975
Beam ($B = 5$)	10.795	33.960	0.109	4.950	8.360	30.612	-0.019	8.950
Beam ($B = 20$)	10.822	34.393	0.116	4.900	8.224	30.537	-0.015	8.900
MBR ($N = 64$)	**15.968**	**43.260**	**0.195**	**4.225**	**11.681**	**37.207**	**0.017**	**6.375**

Table 10: Evaluation of decoding algorithms on speech translation. ∗BLEU scores are used as the utility function for MBR decoding, which may lead to artificially inflated scores that do not accurately reflect a model's true capabilities (Kovacs et al., 2024).

words in Thai (Zeman et al., 2017).[6] Table 8 shows the result. Overall, we observe MBR decoding to consistently outperform beam search in all the languages. The result indicates that the method is effective across different languages.

Comparison to reranking decoding algorithms. In contrast to MBR decoding which selects the hypothesis that has the highest agreement with the other hypotheses, reranking algorithms rescore a fixed set of hypotheses using an external scoring model. We evaluate two reranking algorithms proposed recently. NoRefER selects the sentence with highest score according to a language model fine-tuned for the ASR reranking task (Yuksel et al., 2023).[7] NoRefER does not use the audio input on reranking and relies solely on the generations.

ProGRes selects the hypothesis using the weighted sum of the two objectives, LLM score and ASR score (Tur et al., 2024). LLM score is the perplexity of the hypothesis given a prompt articulated for the reranking task as a context c. We use the same prompt as in Section 2.1 of Tur et al. (2024). Tur et al. (2024) evaluate ProGRes using Llama-3, GPT-3.5, GPT-4 and show that GPT-4 achieves the best performance over the three. Unfortunately, the logits of GPT-3.5 and GPT-4 are no longer provided to the users, so it is not reproducible using these proprietary models. To this end, we use Llama-3 for computing the LLM score in the following experiment (Grattafiori et al., 2024).[8] ASR score is the loss value of the ASR model. We use cross-entropy loss, one of the standard loss functions for ASR models, as the loss function for Whisper is not disclosed (Radford et al., 2023). We set the weight of the LLM score to $\alpha = 0.05$ as it performs the best in the experiments by Tur et al. (2024).

[6] https://pypi.org/project/spacy-thai/
[7] https://huggingface.co/aixplain/NoRefER
[8] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

Table 9 shows the comparison of the reranking algorithms on LibriSpeech and ReazonSpeech. Overall, we observe the performance of the algorithms to be suboptimal compared to MBR decoding and beam search. NoRefER is trained to distinguish models compressed into different sizes so that they have sufficiently different accuracy (Yuksel et al., 2023). Thus, it may be less effective for reranking samples from the same model.

The Oracle score is the score of the hypothesis with the lowest WER (CER) to the reference in the 64 hypotheses sampled. Given that it achieves significantly better score than any of the reranking algorithms, the hypotheses set has good enough hypothesis to be selected and the reranking algorithms have room of improvement.

4.2 Speech Translation

We use the English and Japanese subsets of CoVoST2 (Wang et al., 2021b) and FLEURS (Conneau et al., 2023) datasets for speech translation. We use Kotoba-Whisper-Bilingual for speech translation system.[9] Kotoba-Whisper-Bilingual is a model fine-tuned on top of the distilled Whisper model and trained on a large amount of bilingual speech translation data. It is one of the state-of-the-art open-source systems for bilingual speech recognition and translation for English and Japanese.

We use BLEU using sacrebleu, ROUGE-L (Lin, 2004), BLEURT (Sellam et al., 2020), and MetricX as the evaluation metrics. The other settings are the same as the ASR. Table 10 shows the results of the experiments. Overall, MBR decoding outperforms beam search in all the metrics in both language pairs and datasets.

Note that MBR decoding tends to achieve a relatively higher score than the others on the utility function used during the decoding process (Freitag et al., 2022), which may be indicative of overfitting. Thus, BLEU scores in Table 10 should be interpreted as references.

5 Conclusions

In this paper, we empirically evaluate the performance of MBR decoding for offline ASR and ST tasks. We compare MBR decoding and beam search on a wide range of scenarios with various models, languages, datasets, noise levels, evaluation metrics, and hyperparameters. The extensive evaluation shows that MBR decoding consistently achieves higher accuracy than beam search in both speech-to-text tasks.

The results indicate that MBR decoding has the potential to improve the state-of-the-art performance of offline speech-to-text tasks. Unlike other approaches that depend on heuristics, MBR decoding has a theoretical guarantee (Ichihara et al., 2025a). We believe that MBR decoding is a promising approach for a wide range of speech-to-text tasks and should be considered as one of the baseline methods to improve the system accuracy.

Broader Impact Statement

MBR decoding incurs a computational complexity of $O(UN^2 + GN)$, substantially higher than the $O(GB)$ cost of beam search. When deployed at scale for large-scale transcription tasks, this increased computational burden translates directly into higher energy consumption and carbon emissions. We therefore advise practitioners to measure the computational footprint before deploying MBR in production. A practical strategy to reduce unnecessary computation is the *doubling trick* Besson & Kaufmann (2018); Jinnai & Ariu (2024). Doubling trick starts with a small number of samples, iteratively double the count, and terminate once the selected hypothesis stabilizes between iterations (e.g., the same hypothesis is selected twice in a row). This approach allows practitioners to dynamically bound the computation while retaining the accuracy benefits of MBR decoding.

[9] `https://huggingface.co/kotoba-tech/kotoba-whisper-bilingual-v1.0`

References

Idris Abdulmumin, Victor Agostinelli, Tanel Alumäe, Antonios Anastasopoulos, Luisa Bentivogli, Ondřej Bojar, Claudia Borg, Fethi Bougares, Roldano Cattoni, Mauro Cettolo, Lizhong Chen, William Chen, Raj Dabre, Yannick Estève, Marcello Federico, Mark Fishel, Marco Gaido, Dávid Javorský, Marek Kasztelnik, Fortuné Kponou, Mateusz Krubiński, Tsz Kin Lam, Danni Liu, Evgeny Matusov, Chandresh Kumar Maurya, John P. McCrae, Salima Mdhaffar, Yasmin Moslem, Kenton Murray, Satoshi Nakamura, Matteo Negri, Jan Niehues, Atul Kr. Ojha, John E. Ortega, Sara Papi, Pavel Pecina, Peter Polák, Piotr Połeć, Ashwin Sankar, Beatrice Savoldi, Nivedita Sethiya, Claytone Sikasote, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Brian Thompson, Marco Turchi, Alex Waibel, Patrick Wilken, Rodolfo Zevallos, Vilém Zouhar, and Maike Züfle. Findings of the IWSLT 2025 evaluation campaign. In Elizabeth Salesky, Marcello Federico, and Antonis Anastasopoulos (eds.), *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pp. 412–481, Vienna, Austria (in-person and online), July 2025. Association for Computational Linguistics. ISBN 979-8-89176-272-5. doi: 10.18653/v1/2025.iwslt-1.44. URL https://aclanthology.org/2025.iwslt-1.44/.

Ibrahim Said Ahmad, Antonios Anastasopoulos, Ondřej Bojar, Claudia Borg, Marine Carpuat, Roldano Cattoni, Mauro Cettolo, William Chen, Qianqian Dong, Marcello Federico, Barry Haddow, Dávid Javorský, Mateusz Krubiński, Tsz Kin Lam, Xutai Ma, Prashant Mathur, Evgeny Matusov, Chandresh Maurya, John McCrae, Kenton Murray, Satoshi Nakamura, Matteo Negri, Jan Niehues, Xing Niu, Atul Kr. Ojha, John Ortega, Sara Papi, Peter Polák, Adam Pospíšil, Pavel Pecina, Elizabeth Salesky, Nivedita Sethiya, Balaram Sarkar, Jiatong Shi, Claytone Sikasote, Matthias Sperber, Sebastian Stüker, Katsuhito Sudoh, Brian Thompson, Alex Waibel, Shinji Watanabe, Patrick Wilken, Petr Zemánek, and Rodolfo Zevallos. FINDINGS OF THE IWSLT 2024 EVALUATION CAMPAIGN. In Elizabeth Salesky, Marcello Federico, and Marine Carpuat (eds.), *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pp. 1–11, Bangkok, Thailand (in-person and online), August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.iwslt-1.1. URL https://aclanthology.org/2024.iwslt-1.1/.

Ramya Akula and Ivan Garibay. Sentence pair embeddings based evaluation metric for abstractive and extractive summarization. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pp. 6009–6017, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.646/.

Ahmed Ali and Steve Renals. Word error rate estimation for speech recognition: e-WER. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 20–24, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-2004. URL https://aclanthology.org/P18-2004/.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pp. 4218–4222, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.520/.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 12449–12460. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf.

Waad Ben Kheder, Josef Jon, André Beyer, Abdel Messaoudi, Rabea Affan, Claude Barras, Maxim Tychonov, and Jean-Luc Gauvain. ALADAN at IWSLT24 low-resource Arabic dialectal speech translation

task. In Elizabeth Salesky, Marcello Federico, and Marine Carpuat (eds.), *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pp. 192–202, Bangkok, Thailand (in-person and online), August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.iwslt-1.25. URL https://aclanthology.org/2024.iwslt-1.25/.

Lilian Besson and Emilie Kaufmann. What doubling tricks can and can't do for multi-armed bandits. *arXiv preprint arXiv:1803.06971*, 2018.

Peter J Bickel and Kjell A Doksum. *Mathematical statistics: basic ideas and selected topics, volumes I-II package*. Chapman and Hall/CRC, 2015. doi: https://doi.org/10.1201/9781315369266.

Jean Carletta. Unleashing the killer corpus: experiences in creating the multi-everything ami meeting corpus. *Language Resources and Evaluation*, 41(2):181–190, 2007. doi: https://doi.org/10.1007/s10579-007-9040-x.

Chen Chen, Yuchen Hu, Chao-Han Huck Yang, Sabato Marco Siniscalchi, Pin-Yu Chen, and Eng-Siong Chng. HyPoradise: An open baseline for generative speech recognition with large language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 31665–31688. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/6492267465a7ac507be1f9fd1174e78d-Paper-Datasets_and_Benchmarks.pdf.

Chen Chen, Ruizhe Li, Yuchen Hu, Sabato Marco Siniscalchi, Pin-Yu Chen, EngSiong Chng, and Chao-Han Huck Yang. It's never too late: Fusing acoustic information into large language models for automatic speech recognition. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=QqjFHyQwtF.

Julius Cheng and Andreas Vlachos. Faster minimum Bayes risk decoding with confidence-based pruning. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12473–12480, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.767. URL https://aclanthology.org/2023.emnlp-main.767/.

Chung-Cheng Chiu, Wei Han, Yu Zhang, Ruoming Pang, Sergey Kishchenko, Patrick Nguyen, Arun Narayanan, Hank Liao, Shuyuan Zhang, Anjuli Kannan, Rohit Prabhavalkar, Zhifeng Chen, Tara Sainath, and Yonghui Wu. A comparison of end-to-end models for long-form speech recognition. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 889–896, 2019. doi: 10.1109/ASRU46091.2019.9003854.

Shih-Hsuan Chiu and Berlin Chen. Innovative bert-based reranking language models for speech recognition. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pp. 266–271, 2021. doi: 10.1109/SLT48900.2021.9383557.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, John Hoffman, Min-Jae Hwang, Hirofumi Inaguma, Christopher Klaiber, Ilia Kulikov, Pengwei Li, Daniel Licht, Jean Maillard, Ruslan Mavlyutov, Alice Rakotoarison, Kaushik Ram Sadagopan, Abinesh Ramakrishnan, Tuan Tran, Guillaume Wenzek, Yilin Yang, Ethan Ye, Ivan Evtimov, Pierre Fernandez, Cynthia Gao, Prangthip Hansanti, Elahe Kalbassi, Amanda Kallet, Artyom Kozhevnikov, Gabriel Mejia Gonzalez, Robin San Roman, Christophe Touret, Corinne Wong, Carleigh Wood, Bokai Yu, Pierre Andrews, Can Balioglu, Peng-Jen Chen, Marta R. Costa-jussà, Maha Elbayad, Hongyu Gong, Francisco Guzmán, Kevin Heffernan, Somya Jain, Justine Kao, Ann Lee, Xutai Ma, Alex Mourachko, Benjamin Peloquin, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Anna Sun, Paden Tomasello, Changhan Wang, Jeff Wang, Skyler Wang, and Mary Williamson. Seamless: Multilingual expressive and streaming speech translation. *arXiv preprint arXiv:2312.05187*, 2023. doi: https://doi.org/10.48550/arXiv.2312.05187.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. FLEURS: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pp. 798–805, 2023. doi: 10.1109/SLT54892.2023.10023141.

Hiroyuki Deguchi, Yusuke Sakai, Hidetaka Kamigaito, Taro Watanabe, Hideki Tanaka, and Masao Utiyama. Centroid-based efficient minimum Bayes risk decoding. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 11009–11018, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.654. URL https://aclanthology.org/2024.findings-acl.654/.

Bryan Eikema and Wilker Aziz. Is MAP decoding all you need? the inadequacy of the mode in neural machine translation. In Donia Scott, Nuria Bel, and Chengqing Zong (eds.), *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 4506–4520, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.398. URL https://aclanthology.org/2020.coling-main.398/.

Bryan Eikema and Wilker Aziz. Sampling-based approximations to minimum Bayes risk decoding for neural machine translation. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 10978–10993, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.754. URL https://aclanthology.org/2022.emnlp-main.754/.

J.G. Fiscus. A post-processing system to yield reduced word error rates: Recognizer Output Voting Error Reduction (ROVER). In *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, pp. 347–354, 1997. doi: 10.1109/ASRU.1997.659110.

Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. High quality rather than high model probability: Minimum Bayes risk decoding with neural metrics. *Transactions of the Association for Computational Linguistics*, 10:811–825, 2022. doi: 10.1162/tacl_a_00491. URL https://aclanthology.org/2022.tacl-1.47/.

Markus Freitag, Behrooz Ghorbani, and Patrick Fernandes. Epsilon sampling rocks: Investigating sampling strategies for minimum Bayes risk decoding for machine translation. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9198–9209, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.617. URL https://aclanthology.org/2023.findings-emnlp.617/.

Sanchit Gandhi, Patrick Von Platen, and Alexander M Rush. Distil-whisper: Robust knowledge distillation via large-scale pseudo labelling. *arXiv preprint arXiv:2311.00430v1*, 2023. doi: https://doi.org/10.48550/arXiv.2311.00430.

V. Goel, S. Kumar, and W. Byrne. Segmental minimum bayes-risk decoding for automatic speech recognition. *IEEE Transactions on Speech and Audio Processing*, 12(3):234–249, 2004. doi: 10.1109/TSA.2004.825678.

Vaibhava Goel and William J Byrne. Minimum Bayes-risk automatic speech recognition. *Computer Speech & Language*, 14(2):115–135, 2000. ISSN 0885-2308. doi: https://doi.org/10.1006/csla.2000.0138. URL https://www.sciencedirect.com/science/article/pii/S0885230800901384.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic,

Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan

Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783v3*, 2024. doi: https://doi.org/10.48550/arXiv.2407.21783.

Alex Graves. Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711v1*, 2012. doi: https://doi.org/10.48550/arXiv.1211.3711.

Jinxi Guo, Tara N. Sainath, and Ron J. Weiss. A spelling correction model for end-to-end speech recognition. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5651–5655, 2019. doi: 10.1109/ICASSP.2019.8683745.

John Hewitt, Christopher Manning, and Percy Liang. Truncation sampling as language model desmoothing. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 3414–3427, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.249. URL `https://aclanthology.org/2022.findings-emnlp.249/`.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=rygGQyrFvH`.

Oleksii Hrinchuk, Mariya Popova, and Boris Ginsburg. Correction of automatic speech recognition with transformer sequence-to-sequence model. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7074–7078, 2020. doi: 10.1109/ICASSP40776.2020.9053051.

Ke Hu, Tara N. Sainath, Ruoming Pang, and Rohit Prabhavalkar. Deliberation model based two-pass end-to-end speech recognition. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7799–7803, 2020. doi: 10.1109/ICASSP40776.2020.9053606.

Yuchen Hu, Chen Chen, Chao-Han Huck Yang, Ruizhe Li, Chao Zhang, Pin-Yu Chen, and Engsiong Chng. Large language models are efficient learners of noise-robust speech recognition. In *ICLR*, 2024. URL `https://openreview.net/forum?id=ceATjGPTUD`.

Yuki Ichihara, Yuu Jinnai, Kaito Ariu, Tetsuro Morimura, and Eiji Uchibe. Theoretical guarantees for minimum Bayes risk decoding. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computa-*

tional Linguistics (Volume 1: Long Papers), pp. 16262–16284, Vienna, Austria, July 2025a. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.793. URL https://aclanthology.org/2025.acl-long.793/.

Yuki Ichihara, Yuu Jinnai, Tetsuro Morimura, Kenshi Abe, Kaito Ariu, Mitsuki Sakamoto, and Eiji Uchibe. Evaluation of best-of-n sampling strategies for language model alignment. *Transactions on Machine Learning Research*, 2025b. ISSN 2835-8856. URL https://openreview.net/forum?id=H4S4ETc8c9.

Yuu Jinnai and Kaito Ariu. Hyperparameter-free approach for faster minimum Bayes risk decoding. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8547–8566, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.505. URL https://aclanthology.org/2024.findings-acl.505/.

Yuu Jinnai, Tetsuro Morimura, Ukyo Honda, Kaito Ariu, and Kenshi Abe. Model-based minimum Bayes risk decoding for text generation. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 22326–22347. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/jinnai24a.html.

Juraj Juraska, Mara Finkelstein, Daniel Deutsch, Aditya Siddhant, Mehdi Mirzazadeh, and Markus Freitag. MetricX-23: The Google submission to the WMT 2023 metrics shared task. In Philipp Koehn, Barry Haddow, Tom Kocmi, and Christof Monz (eds.), *Proceedings of the Eighth Conference on Machine Translation*, pp. 756–767, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.wmt-1.63. URL https://aclanthology.org/2023.wmt-1.63/.

Kentaro Kamiya, Takuya Kawase, Ryuichiro Higashinaka, and Katashi Nagao. Using presentation slides and adjacent utterances for post-editing of speech recognition results for meeting recordings. In Kamil Ekštein, František Pártl, and Miloslav Konopík (eds.), *Text, Speech, and Dialogue*, pp. 331–340, Cham, 2021. Springer International Publishing. ISBN 978-3-030-83527-9. doi: https://doi.org/10.1007/978-3-030-83527-9_28.

Naoyuki Kamo, Tsubasa Ochiai, Marc Delcroix, and Tomohiro Nakatani. MOVER: Combining Multiple Meeting Recognition Systems. In *Interspeech 2025*, pp. 3424–3428, 2025. doi: 10.21437/Interspeech.2025-1614.

Suyoun Kim, Abhinav Arora, Duc Le, Ching-Feng Yeh, Christian Fuegen, Ozlem Kalinli, and Michael L. Seltzer. Semantic distance: A new metric for ASR performance analysis towards spoken language understanding. In *Interspeech 2021*, pp. 1977–1981, 2021. doi: 10.21437/Interspeech.2021-1929.

Nithin Rao Koluguri, Samuel Kriman, Georgy Zelenfroind, Somshubra Majumdar, Dima Rekesh, Vahid Noroozi, Jagadeesh Balam, and Boris Ginsburg. Investigating end-to-end ASR architectures for long form audio transcription. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 13366–13370, 2024. doi: 10.1109/ICASSP48485.2024.10448309.

Geza Kovacs, Daniel Deutsch, and Markus Freitag. Mitigating metric bias in minimum Bayes risk decoding. In Barry Haddow, Tom Kocmi, Philipp Koehn, and Christof Monz (eds.), *Proceedings of the Ninth Conference on Machine Translation*, pp. 1063–1094, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.wmt-1.109. URL https://aclanthology.org/2024.wmt-1.109/.

Taku Kudo. Mecab: Yet another part-of-speech and morphological analyzer, 2005. URL https://taku910.github.io/mecab.

Shankar Kumar and William Byrne. Minimum Bayes-risk decoding for statistical machine translation. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pp. 169–176, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for Computational Linguistics. URL https://aclanthology.org/N04-1022/.

Zhihong Lei, Mingbin Xu, Shiyi Han, Leo Liu, Zhen Huang, Tim Ng, Yuanyuan Zhang, Ernest Pusateri, Mirko Hannemann, Yaqiao Deng, and Man-Hung Siu. Acoustic model fusion for end-to-end speech recognition. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–7, 2023. doi: 10.1109/ASRU57964.2023.10389720.

Yichong Leng, Xu Tan, Linchen Zhu, Jin Xu, Renqian Luo, Linquan Liu, Tao Qin, Xiangyang Li, Edward Lin, and Tie-Yan Liu. FastCorrect: Fast error correction with edit alignment for automatic speech recognition. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 21708–21719. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/b597460c506e8e35fb0cc1c1905dd3bc-Paper.pdf.

Zhaolin Li, Enes Yavuz Ugan, Danni Liu, Carlos Mullov, Tu Anh Dinh, Sai Koneru, Alexander Waibel, and Jan Niehues. The KIT speech translation systems for IWSLT 2024 dialectal and low-resource track. In Elizabeth Salesky, Marcello Federico, and Marine Carpuat (eds.), *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pp. 221–228, Bangkok, Thailand (in-person and online), August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.iwslt-1.27. URL https://aclanthology.org/2024.iwslt-1.27/.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pp. 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/W04-1013/.

Danni Liu, Jan Niehues, and Gerasimos Spanakis. Adapting end-to-end speech recognition for readable subtitles. In Marcello Federico, Alex Waibel, Kevin Knight, Satoshi Nakamura, Hermann Ney, Jan Niehues, Sebastian Stüker, Dekai Wu, Joseph Mariani, and Francois Yvon (eds.), *Proceedings of the 17th International Conference on Spoken Language Translation*, pp. 247–256, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.iwslt-1.30. URL https://aclanthology.org/2020.iwslt-1.30/.

Rao Ma, Mengjie Qian, Potsawee Manakul, Mark Gales, and Kate Knill. Can generative large language models perform ASR error correction? *arXiv preprint arXiv:2307.04172v2*, 2023. doi: https://doi.org/10.48550/arXiv.2307.04172.

Fabrizio Morbini, Kartik Audhkhasi, Ron Artstein, Maarten Van Segbroeck, Kenji Sagae, Panayiotis Georgiou, David R. Traum, and Shri Narayanan. A reranking approach for recognition and classification of speech input in conversational dialogue systems. In *2012 IEEE Spoken Language Technology Workshop (SLT)*, pp. 49–54, 2012. doi: 10.1109/SLT.2012.6424196.

Ananya Mukherjee and Manish Shrivastava. Unsupervised embedding-based metric for MT evaluation with improved human correlation. In Philipp Koehn, Loïc Barrault, Ondřej Bojar, Fethi Bougares, Rajen Chatterjee, Marta R. Costa-jussà, Christian Federmann, Mark Fishel, Alexander Fraser, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Paco Guzman, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Tom Kocmi, André Martins, Makoto Morishita, Christof Monz, Masaaki Nagata, Toshiaki Nakazawa, Matteo Negri, Aurélie Névéol, Mariana Neves, Martin Popel, Marco Turchi, and Marcos Zampieri (eds.), *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pp. 558–563, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics. URL https://aclanthology.org/2022.wmt-1.49/.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2022.

Arun Narayanan, Rohit Prabhavalkar, Chung-Cheng Chiu, David Rybach, Tara N. Sainath, and Trevor Strohman. Recognizing long-form speech using streaming end-to-end models. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 920–927, 2019. doi: 10.1109/ASRU46091.2019.9003913.

Matteo Negri, Marco Turchi, José G. C. de Souza, and Daniele Falavigna. Quality estimation for automatic speech recognition. In Junichi Tsujii and Jan Hajic (eds.), *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pp. 1813–1823, Dublin, Ireland, August 2014. Dublin City University and Association for Computational Linguistics. URL https://aclanthology.org/C14-1171/.

Raymond W. M. Ng, Kashif Shah, Wilker Aziz, Lucia Specia, and Thomas Hain. Quality estimation for ASR k-best list rescoring in spoken language translation. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5226–5230, 2015. doi: 10.1109/ICASSP.2015.7178968.

Mengxi Nie, Ming Yan, and Caixia Gong. Prompt-based re-ranking language model for asr. In *Interspeech 2022*, pp. 3864–3868, 2022. doi: 10.21437/Interspeech.2022-536.

Atsumoto Ohashi, Ukyo Honda, Tetsuro Morimura, and Yuu Jinnai. On the true distribution approximation of minimum Bayes-risk decoding. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 459–468, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.naacl-short.38. URL https://aclanthology.org/2024.naacl-short.38/.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. LibriSpeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5206–5210, 2015. doi: 10.1109/ICASSP.2015.7178964.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin (eds.), *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040/.

Aditya Kamlesh Parikh, Louis ten Bosch, and Henk van den Heuvel. Ensembles of hybrid and end-to-end speech recognition. In Nicoletta Calzolari, Min-Yen Kan, Veronique Hoste, Alessandro Lenci, Sakriani Sakti, and Nianwen Xue (eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 6199–6205, Torino, Italia, May 2024. ELRA and ICCL. URL https://aclanthology.org/2024.lrec-main.547/.

Matt Post. A call for clarity in reporting BLEU scores. In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor (eds.), *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 186–191, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6319. URL https://aclanthology.org/W18-6319/.

Rohit Prabhavalkar, Takaaki Hori, Tara N. Sainath, Ralf Schlüter, and Shinji Watanabe. End-to-end speech recognition: A survey. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:325–351, 2024. doi: 10.1109/TASLP.2023.3328283.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 28492–28518. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/radford23a.html.

Srijith Radhakrishnan, Chao-Han Huck Yang, Sumeer Ahmad Khan, Rohit Kumar, Narsis A. Kiani, David Gomez-Cabrero, and Jesper N. Tegner. Whispering LLaMA: A cross-modal generative error correction framework for speech recognition. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 10007–10016, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.618. URL https://aclanthology.org/2023.emnlp-main.618/.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 3982–3992, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1410. URL https://aclanthology.org/D19-1410/.

Nils Reimers and Iryna Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4512–4525, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.365. URL https://aclanthology.org/2020.emnlp-main.365/.

Nathaniel Romney Robinson, Niyati Bafna, Xiluo He, Tom Lupicki, Lavanya Shankar, Cihan Xiao, Qi Sun, Kenton Murray, and David Yarowsky. JHU IWSLT 2025 low-resource system description. In Elizabeth Salesky, Marcello Federico, and Antonis Anastasopoulos (eds.), *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pp. 315–323, Vienna, Austria (in-person and online), July 2025. Association for Computational Linguistics. ISBN 979-8-89176-272-5. doi: 10.18653/v1/2025.iwslt-1.32. URL https://aclanthology.org/2025.iwslt-1.32/.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. Masked language model scoring. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2699–2712, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.240. URL https://aclanthology.org/2020.acl-main.240/.

Toshinori Sato, Taiichi Hashimoto, and Manabu Okumura. Implementation of a word segmentation dictionary called mecab-ipadic-neologd and study on how to use it effectively for information retrieval (in japanese). In *Proceedings of the Twenty-three Annual Meeting of the Association for Natural Language Processing*, pp. NLP2017–B6–1. The Association for Natural Language Processing, 2017. URL https://www.anlp.jp/proceedings/annual_meeting/2017/pdf_dir/B6-1.pdf.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. BLEURT: Learning robust metrics for text generation. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault (eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7881–7892, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.704. URL https://aclanthology.org/2020.acl-main.704/.

Prashanth Gurunath Shivakumar, Jari Kolehmainen, Aditya Gourav, Yi Gu, Ankur Gandhe, Ariya Rastrow, and Ivan Bulyko. Speech recognition rescoring with large speech-text foundation models. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2025. doi: 10.1109/ICASSP49660.2025.10890616.

David Snyder, Guoguo Chen, and Daniel Povey. Musan: A music, speech, and noise corpus. *arXiv preprint arXiv:1510.08484v1*, 2015. doi: https://doi.org/10.48550/arXiv.1510.08484.

Ryosuke Sonobe, Shinnosuke Takamichi, and Hiroshi Saruwatari. JSUT corpus: free large-scale japanese speech corpus for end-to-end speech synthesis. *arXiv preprint arXiv:1711.00354v1*, 2017. doi: https://doi.org/10.48550/arXiv.1711.00354.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper_files/paper/2014/file/5a18e133cbf9f257297f410bb7eca942-Paper.pdf.

Mirac Suzgun, Luke Melas-Kyriazi, and Dan Jurafsky. Follow the wisdom of the crowd: Effective text generation via minimum Bayes risk decoding. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 4265–4293, Toronto,

Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.262. URL https://aclanthology.org/2023.findings-acl.262/.

Conghui Tan, Di Jiang, Jinhua Peng, Xueyang Wu, Qian Xu, and Qiang Yang. A de novo divide-and-merge paradigm for acoustic model optimization in automatic speech recognition. In Christian Bessiere (ed.), *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pp. 3709–3715. International Joint Conferences on Artificial Intelligence Organization, 7 2020. doi: 10.24963/ijcai.2020/513. URL https://doi.org/10.24963/ijcai.2020/513. Main track.

Firas Trabelsi, David Vilar, Mara Finkelstein, and Markus Freitag. Efficient minimum bayes risk decoding using low-rank matrix completion algorithms. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 54714–54733. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/626ab938fe19200324b368f5ee816868-Paper-Conference.pdf.

Ada Defne Tur, Adel Moumen, and Mirco Ravanelli. ProGRes: Prompted generative rescoring on ASR N-best. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pp. 600–607, 2024. doi: 10.1109/SLT61566.2024.10832194.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Abdul Waheed, Hanin Atwany, Rita Singh, and Bhiksha Raj. On the robust approximation of ASR metrics. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 23119–23146, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.1187. URL https://aclanthology.org/2025.findings-acl.1187/.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. Fairseq S2T: Fast speech-to-text modeling with fairseq. In Derek Wong and Douwe Kiela (eds.), *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations*, pp. 33–39, Suzhou, China, December 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.aacl-demo.6. URL https://aclanthology.org/2020.aacl-demo.6/.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 993–1003, Online, August 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.80. URL https://aclanthology.org/2021.acl-long.80/.

Changhan Wang, Anne Wu, Jiatao Gu, and Juan Pino. CoVoST 2 and massively multilingual speech translation. In *Interspeech 2021*, pp. 2247–2251, 2021b. doi: 10.21437/Interspeech.2021-2027.

Wenxuan Wang, Yingxin Zhang, Yifan Jin, Binbin Du, and Yuke Li. NYA's offline speech translation system for IWSLT 2025. In Elizabeth Salesky, Marcello Federico, and Antonis Anastasopoulos (eds.), *Proceedings of the 22nd International Conference on Spoken Language Translation (IWSLT 2025)*, pp. 206–211, Vienna, Austria (in-person and online), July 2025. Association for Computational Linguistics. ISBN 979-8-89176-272-5. doi: 10.18653/v1/2025.iwslt-1.19. URL https://aclanthology.org/2025.iwslt-1.19/.

Sam Wiseman, Stuart Shieber, and Alexander Rush. Challenges in data-to-document generation. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel (eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 2253–2263, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/D17-1239. URL https://aclanthology.org/D17-1239/.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. Transformers: State-of-the-art natural language processing. In Qun Liu and David Schlangen (eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 38–45, Online, October 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-demos.6. URL https://aclanthology.org/2020.emnlp-demos.6/.

Ian Wu, Patrick Fernandes, Amanda Bertsch, Seungone Kim, Sina Khoshfetrat Pakazad, and Graham Neubig. Better Instruction-Following Through Minimum Bayes Risk. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=7xCSK9BLPy.

Haihua Xu, Daniel Povey, Lidia Mangu, and Jie Zhu. An improved consensus-like method for minimum Bayes risk decoding and lattice combination. In *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 4938–4941, 2010. doi: 10.1109/ICASSP.2010.5495100.

Haihua Xu, Daniel Povey, Lidia Mangu, and Jie Zhu. Minimum Bayes risk decoding and system combination based on a recursion for edit distance. *Computer Speech & Language*, 25(4):802–828, 2011. ISSN 0885-2308. doi: https://doi.org/10.1016/j.csl.2011.03.001. URL https://www.sciencedirect.com/science/article/pii/S0885230811000192.

Liyan Xu, Yile Gu, Jari Kolehmainen, Haidar Khan, Ankur Gandhe, Ariya Rastrow, Andreas Stolcke, and Ivan Bulyko. RescoreBERT: Discriminative speech recognition rescoring with BERT. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6117–6121, 2022. doi: 10.1109/ICASSP43922.2022.9747118.

Brian Yan, Patrick Fernandes, Jinchuan Tian, Siqi Ouyang, William Chen, Karen Livescu, Lei Li, Graham Neubig, and Shinji Watanabe. CMU's IWSLT 2024 offline speech translation system: A cascaded approach for long-form robustness. In Elizabeth Salesky, Marcello Federico, and Marine Carpuat (eds.), *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pp. 164–169, Bangkok, Thailand (in-person and online), August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.iwslt-1.22. URL https://aclanthology.org/2024.iwslt-1.22/.

Chao-Han Huck Yang, Yile Gu, Yi-Chieh Liu, Shalini Ghosh, Ivan Bulyko, and Andreas Stolcke. Generative speech recognition error correction with large language models and task-activating prompting. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–8, 2023. doi: 10.1109/ASRU57964.2023.10389673.

Yue Yin, Daijiro Mori, and Seiji Fujimoto. ReazonSpeech: A free and massive corpus for japanese ASR (in japanese). In *Proceedings of the Annual Meeting of the Association for Natural Language Processing*, pp. 1134–1139, 2023. URL https://www.anlp.jp/proceedings/annual_meeting/2023/pdf_dir/A5-3.pdf.

Kamer Ali Yuksel, Thiago Ferreira, Ahmet Gunduz, Mohamed Al-Badrashiny, and Golara Javadi. A reference-less quality metric for automatic speech recognition via contrastive-learning of a multi-language model with self-supervision. In *2023 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pp. 1–5, 2023. doi: 10.1109/ICASSPW59220.2023.10193003.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy,

Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria de Paiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonça, Tatiana Lando, Rattima Nitisaroj, and Josie Li. CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies. In Jan Hajič and Dan Zeman (eds.), *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pp. 1–19, Vancouver, Canada, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/K17-3001. URL `https://aclanthology.org/K17-3001/`.

A Limitations

One of the critical limitations of MBR decoding is the computational cost. For the sake of reference, we provide the walltime of the decoding algorithms with our implementation (Appendix B).

There are several libraries dedicated to optimizing the speed of the whisper models, such as faster-whisper[10] and whisper.cpp.[11] Thus, beam search can be made faster by using such libraries, but MBR decoding may require additional modifications to fully leverage these optimizations. Developing a fast implementation of MBR decoding is left for future work.

Experiments are primarily conducted using sequence-to-sequence autoregressive models (Vaswani et al., 2017; Radford et al., 2023). We also attempted to apply MBR decoding to Wav2Vec 2.0 (Baevski et al., 2020), a CTC-based model. CTC models produce extremely peaked per-frame probability distributions: random sampling yields the same result as greedy (MAP) decoding unless the temperature is raised to a level that severely degrades output quality. Consequently, MBR decoding is not directly applicable to CTC-based models in their standard configuration. Evaluation of MBR decoding to wider range of models remains an open direction for future work.

The Musan dataset (Snyder et al., 2015) covers a wide range of noise types, but it may not fully represent the noise encountered in all the communities and regions. Evaluation using real-world noisy datasets for the particular communities and regions is left for future work.

B Walltime

Table 11 shows the average walltime on the LibriSpeech Clean dataset with the whisper-large-v3 model. Note that because the experiment is not conducted to evaluate the walltime of the decoding algorithms, our codebase is not optimized to reduce the walltime. For example, the reported values include the time for logging and sending the generated hypotheses to a cloud server, which adds to the overall time. Also note that the walltime also depends on the choice of the utility function. Currently, computing the BLEU scores on CPU is taking the majority of the computation time. We find that using SentBERT as the utility function is much faster than using BLEU, as SentBERT runs on a GPU in parallel and does not require CPU/GPU data transfer. Thus, the reported time does not reflect the performance of optimized implementations and should solely be considered as a reference.

C Reproducibility Statement

The code will be available as an open source upon publication. All the experiments are conducted using publicly available resources shown in Table 12.

[10] `https://github.com/systran/faster-whisper`
[11] `https://github.com/ggml-org/whisper.cpp`

Method	Walltime (seconds)	WER
Beam ($B = 1$)	0.88	0.042
Beam ($B = 5$)	1.54	0.042
Beam ($B = 20$)	1.56	0.042
MBR ($N = 4$)	2.47	0.035
MBR ($N = 8$)	3.44	0.035
MBR ($N = 16$)	7.97	0.034
MBR ($N = 32$)	17.89	0.032
MBR ($N = 64$)	30.18	0.033

Table 11: Estimated average walltime of the decoding algorithms on the LibriSpeech dataset with the Whisper-large-v3 model. Note that the walltime includes the time for logging and sending the generated hypotheses to a cloud server for record, which adds to the overall time. Thus, the reported time does not reflect the performance of optimized implementations and should solely be considered as a reference.

Datasets	
LibriSpeech	https://huggingface.co/datasets/openslr/librispeech_asr (Panayotov et al., 2015)
VoxPopuli	https://huggingface.co/datasets/facebook/voxpopuli (Wang et al., 2021a)
AMI-IHM	https://huggingface.co/datasets/edinburghcstr/ami (Carletta, 2007)
ReazonSpeech	https://huggingface.co/datasets/japanese-asr/ja_asr.reazonspeech_test (Yin et al., 2023)
CommonVoice-v8	https://huggingface.co/datasets/mozilla-foundation/common_voice_8_0 (Ardila et al., 2020)
JSUT	https://huggingface.co/datasets/japanese-asr/ja_asr.jsut_basic5000 (Sonobe et al., 2017)
CoVoST2	https://huggingface.co/datasets/facebook/covost2 (Wang et al., 2021b)
FLEURS	https://huggingface.co/datasets/google/fleurs (Conneau et al., 2023)
Models	
whisper-large-v3	https://huggingface.co/openai/whisper-large-v3 (Radford et al., 2023)
whisper-small	https://huggingface.co/openai/whisper-small (Radford et al., 2023)
whisper-medium	https://huggingface.co/openai/whisper-medium (Radford et al., 2023)
distil-whisper	https://huggingface.co/distil-whisper/distil-large-v3.5 (Gandhi et al., 2023)
kotoba-whisper	https://huggingface.co/kotoba-tech/kotoba-whisper-v2.0
kotoba-whisper-bilingual	https://huggingface.co/kotoba-tech/kotoba-whisper-bilingual-v1.0
Others	
BLEURT	https://huggingface.co/lucadiliello/BLEURT-20-D12 (Sellam et al., 2020)
MetricX	https://huggingface.co/google/metricx-23-xxl-v2p0 (Juraska et al., 2023)
all-MiniLM-L6-v2	https://huggingface.co/sentence-transformers/all-mpnet-base-v2 (Reimers & Gurevych, 2019; 2020)
NoRefER	Because only part of the code is published (https://huggingface.co/aixplain/NoRefER), the method is implemented by us. (Yuksel et al., 2023)
ProGRes	Because only part of the code is published (https://github.com/AdaDTur/ProGRes), the method is implemented by us. (Tur et al., 2024)
Llama-3	https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct (Grattafiori et al., 2024)

Table 12: List of datasets and models used in this study. All the resources are publicly available.