Mai Hoʻomāuna i ka ʻAi: Language Models Improve Automatic Speech Recognition in Hawaiian

Kaavya Chaparala
The MITRE Corporation
kchaparala@mitre.org

Guido Zarrella
The MITRE Corporation
jzarrella@mitre.org

Bruce Torres Fischer
University of Hawaiʻi at Hilo
bruce42@hawaii.edu

Larry Kimura
University of Hawaiʻi at Hilo
larrykim@hawaii.edu

Oiwi Parker Jones
University of Oxford
oiwi@robots.ox.ac.uk

Abstract

In this paper we address the challenge of improving Automatic Speech Recognition (ASR) for a low-resource language, Hawaiian, by incorporating large amounts of independent text data into an ASR foundation model, Whisper. To do this, we train an external language model (LM) on $\sim$1.5M words of Hawaiian text. We then use the LM to rescore Whisper and compute word error rates (WERs) on a manually curated test set of labeled Hawaiian data. As a baseline, we use Whisper without an external LM. Experimental results reveal a small but significant improvement in WER when ASR outputs are rescored with a Hawaiian LM. The results support leveraging all available data in the development of ASR systems for underrepresented languages.

1 Introduction

Advances in Automatic Speech Recognition (ASR) have predominantly benefited a few global languages. To date, this has left speakers of low-resource languages, such as Hawaiian, at a disadvantage. High-quality Hawaiian ASR would support language preservation and revitalization efforts. In Hawaiian there is a traditional saying – *Mai hoʻomāuna i ka ʻai* – which translates literally as ‘Don’t waste food’ (1). It can also be applied in the context of ASR to mean ‘Leave nothing to waste’ or even ‘Use all the data you have’. In this paper, we explore the use of text corpora as a potentially underutilized form of data to improve low-resource ASR systems. We do this by comparing a foundation model, Whisper (2), both with and without the use of a relatively large amount of independent Hawaiian text. To leverage these data, we use an external language model (LM).

Unlike other early ASR foundation models (e.g. wav2vec 2.0 (3) and XLS-R (4)), Whisper is explicitly intended for *zero-shot transfer* (2). This means that the model is trained on a large and diverse set of data to generalize well on new ASR tasks without fine-tuning. The first contribution of this paper is to quantify Whisper’s ability to perform zero-shot transfer in Hawaiian. We quantify performance using word error rates (WERs) calculated on a manually curated test set of labeled data (i.e. Hawaiian audio–text pairs).

Prior work has explored how to improve foundation models like Whisper by incorporating independent large language models (LLMs) for English (5). To the best of our knowledge, no similar work has been reported for low-resources languages like Hawaiian. Therefore, the second contribution of this paper is to evaluate the utility of incorporating independent Hawaiian text data into Whisper. We do this by replicating a state-of-the-art Hawaiian LM (6; 7) which we train on a relatively large corpus of modern Hawaiian text ($\sim$1.5M words). We then compare a baseline zero-shot Whisper model against a zero-shot Whisper model that has been rescored (8) using the Hawaiian LM.

4th NeurIPS Efficient Natural Language and Speech Processing Workshop (ENLSP-IV 2024).

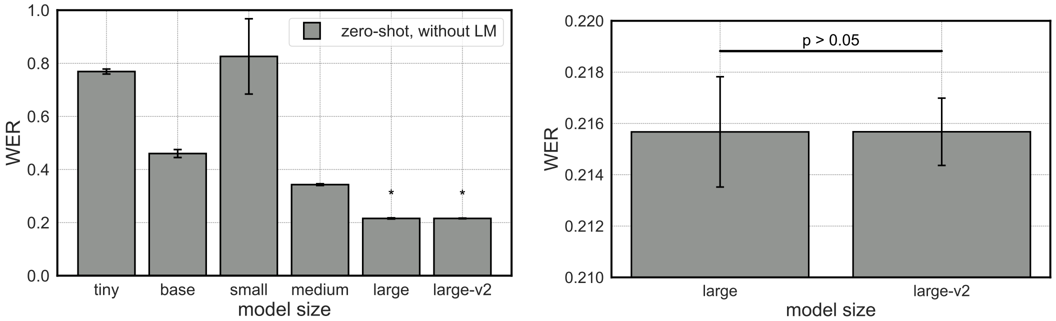

Figure 1: **Large ASR models produce the lowest word error rates (WERs) for Hawaiian test data.** Left panel: We compared six Whisper models on Hawaiian using zero-shot transfer without a Hawaiian language model (LM), as a baseline for comparing ASR models with LMs. Asterisks indicate the best models, `large` and `large-v2`. Right panel: No statistical difference in WER was observed between `large` and `large-v2` ($t_{3.309} = 0.002, p = 0.999$, Welch's t-test). Error bars show standard error of the mean.

2 Methods

2.1 Hawaiian LMs

Here we replicate the model architecture and training parameters from prior work on Hawaiian LMs (6; 7), using the same dataset and data splits. Concretely, the training set, which we used for training LMs, alternately consists of 45,769 lines, 1,547,831 words, or 7,573,569 characters. The validation set, which we used for evaluating the validation loss and perplexity, consists of 888 lines, 26,607 words, or 129,487 characters. We did not use the test set (888 lines, 30,181 words, 146,124 characters) for this project. All modern Hawaiian texts (e.g. (9)) were used with permission.

For the model architecture we followed prior work (6; 7) in implementing a character-level RNN (10; 11) which consisted of three layers of LSTMs (12) each with 200 features in their hidden states. The output of the final LSTM layer was passed to a linear layer that mapped to the set of output characters. We applied Dropout (13) with a probability of 0.2 after the input and after each hidden layer. For the loss, we used cross entropy. LMs were trained using Adam (14) for 10,000 epochs with an initial learning rate of 0.001, a batch size of 256, a clip value of 1, and a maximum sequence length of 100 characters. During training, we checkpointed the LM that produced the lowest perplexity on the validation set. In the final LM which we used for rescoring, the validation perplexity was 2.024. This is, incidentally, better than the reported state-of-the-art validation perplexity of 2.65 (7), a difference that we attribute to our checkpointing strategy.

For rescoring, the LM was used to compute the log probability of strings. This computation involved prepending a start-of-sequence (SOS) token to the string. The LM needs this to transition into the first character in the string. Since the LM character set did not include an explicit SOS token, we opted to use a whitespace character instead. This decision was based on the rationale that, within the context of the LM, the probability distribution following a whitespace character should closely approximate the distribution of initial characters in words.

Formally, the log probability of observing a character y_i, given the LM and preceding sequence of characters, is denoted as $\log P_{\text{LM}}(y_i|y_{1:i-1})$, where $y_{1:i-1}$ represents the sequence of characters from the start of the string up to, but not including, y_i. Consequently, the log probability of an entire string $Y = y_1, y_2, \ldots, y_n$ is computed as the sum of its individual characters, given by their respective preceding character sequences: $\log P_{\text{LM}}(Y) = \sum_{i=1}^{n} \log P_{\text{LM}}(y_i|y_{1:i-1})$.

2.2 Rescoring Whisper

Encoder–decoder models, such as Whisper (2), are essentially conditional language models that learn an implicit language model over the domain of output tokens in their training data. Critically, Whisper's implicit language model is not tuned to Hawaiian, as no Hawaiian text was included during training. We note that some Hawaiian audio was paired with English text in Whisper's $X \rightarrow \texttt{en}$ task (2). But Whisper was not trained on any Hawaiian audio–text pairs, or indeed any Hawaiian text. We therefore propose to incorporate an explicit Hawaiian LM into Whisper via rescoring. In rescoring

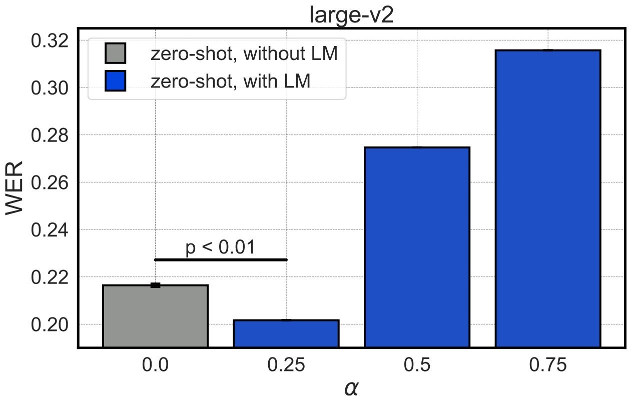

Figure 2: **Rescoring with a Hawaiian LM provides a small but significant improvement on the zero-shot Whisper baseline.** Rescoring results for `large-v2`. The α values weight the contribution of the LM. $\alpha = 0$ means no contribution of the LM (baseline model). Other values add increasing weight to the LM. The best WER was found at $\alpha = 0.25$ where we observe a small but significant improvement on the baseline model ($t_2 = 19.498, p = 0.003$, one-sample t-test).

(8), beam search utilises a weighted sum of the ASR model's log probability of a given hypothesis and the external LM's log probability of the same hypothesis.

To see this, let the ASR model's log probability be $\log P_{\text{ASR}}(Y|X)$, where Y is the output text and X is the input audio. We combine this with the log probability of the external LM, $\log P_{\text{LM}}(Y)$, as follows:

$$\text{score} = \alpha \log P_{\text{LM}}(Y) + (1 - \alpha) \log P_{\text{ASR}}(Y|X) \tag{1}$$

Here α is a hyperparameter that balances the influence of the log probabilities of the LM and ASR models. As expressed, $\alpha = 0$ means no contribution from the LM. We explore values of $\alpha \in \{0, 0.25, 0.5, 0.75\}$. We exclude $\alpha = 1$ as this would mean no influence from the ASR.

In the Hugging Face implementation that we used,[1] Whisper employs a multi-stage decoding process. It begins with beam search, utilizing a default beam size of 5. However, if the output does not meet predefined compression ratio and log-probability thresholds, the implementation iteratively resorts to greedy search with temperatures varying between 0.2 to 1.0. These heuristics aim to ensure that the final hypothesis avoids common error modes for encoder–decoder models, such as repeating substrings in the output or failing to generate an end-of-text token. We highlight these heuristics as they affect whether model outputs are deterministic or stochastic. This in turn affected our choice of statistical test when comparing different Whisper models (one sample vs independent samples t-tests).

2.3 ASR Test Set

The evaluation of WERs required a labeled ASR test set for Hawaiian. To this end, we utilized a subset of the publicly available *Ka Leo Hawai'i* (KLH) dataset, being careful to account for transcription and segmentation errors within the data. First, we randomly selected 100 audio–text pairs from the dataset. We then filtered the pairs manually, discarding any that, after text normalization, resulted in empty strings (e.g. from audio segments annotating only laughter). We also discarded pairs that showed discrepancies between audio and text (e.g. truncated recordings). Such exclusions were crucial to avoid artificially inflating the WER through misalignments. After filtering 57 pairs remained, comprising 1,120 words and a total audio duration of 7 minutes and 35.336 seconds. This constituted our final test set. As Whisper does not process audio files longer than 30 seconds, it is important to note that the longest audio file in the ASR test set was only 26.593 seconds in duration.

[1] `https://huggingface.co/docs/transformers/en/model_doc/whisper`

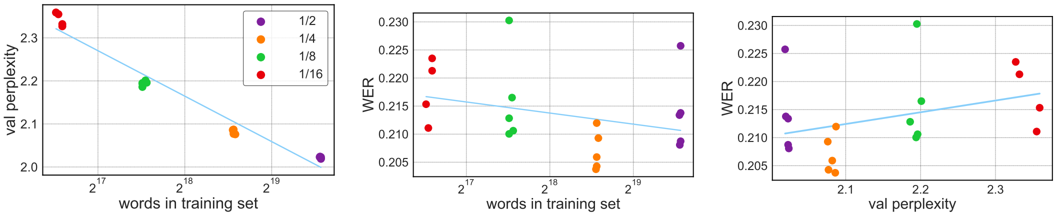

Figure 3: **Posthoc exploration on the amount of training text, LM validation perplexity, and Whisper WER.** Hawaiian LMs were trained on decreasing fractions of data: 1/2 (purple), 1/4 (orange), 1/8 (green), and 1/16 (red). See text for details.

3 Experiment Results

3.1 Which Whisper model transfers best to Hawaiian?

Whisper comes in different sizes: `tiny`, `base`, `small`, `medium`, and `large`. Multiple `large` models were also available when we ran the experiments: `large` and `large-v2`. We compared all of these using zero-shot transfer on the ASR test for Hawaiian, repeating the evaluation three times for each model to estimate the variability in WER results. The best results (Figure 1, left panel) were obtained for the `large` and `large-v2` models. Focusing on the best two models (Figure 1, right panel), we find no significant difference between `large` and `large-v2` in terms of zero-shot WER ($t_{3.309} = 0.002, p = 0.999$, Welch's t-test). As either model could be used, we focus below on rescoring `large-v2`.

3.2 Does rescoring improve the best Whisper model?

Figure 2 shows that rescoring the `large-v2` model with the Hawaiian LM produces a small absolute improvement (~1–2%) in WER when compared to the zero-shot baseline at $\alpha = 0.25$. This improvement is, however, significant ($t_2 = 19.498, p = 0.003$, one-sample t-test). At the time of writing, it also represents the state-of-the-art in Hawaiian ASR. To get a sense of what predictions from the model look like, we can sample a few examples at random. Predicted transcripts from the best rescored zero-shot model are compared against ground truth transcripts in Table 1.

3.3 Does it matter how much text the LM is trained on?

In a final set of posthoc experiments, we re-trained a series of Hawaiian language models on fractions of the full dataset used in the main results. Starting with the full training set (45,769 lines), we randomly sampled subsets of training text with varying fractions of lines: 1/2, 1/4, 1/8, and 1/16. For each fraction of lines, we created five randomly sub-sampled training sets. These were then used to train 20 new LMs (4 fractions × 5 repeats) using the same training parameters as in the main analysis. These 20 LMs were then used to rescore the `large-v2` Whisper model.

In Figure 3, we show scatter plots that compare the number of words in the sub-sampled training set, the trained LM's validation perplexity, and the `large-v2` Whisper model's rescored WER.

We observe a negative correlation between the number of words in the training set and the LM's validation perplexity ($r = -0.98, p = 8.67 \times 10^{-13}$, Pearson's correlation, Bonferroni corrected, 3 comparisons). More text for training corresponds to smaller validation perplexity (Figure 3, left panel).

Correlations involving WER did not reach significance, apparently because of large variance in the WER results. However, linear regression models suggest trends. For example, although not statistically significant ($r = -0.308, p = 0.14$ uncorrected), the trend is for WER to decrease as training data increases (Figure 3, middle panel). Similarly, although not statistically significant ($r = 0.352, p = 0.14$ uncorrected), WER appears to increase as validation perplexity increases (Figure 3, right panel). Or, equivalently, in this last case, WER appears to decrease as validation perplexity decreases. Again, we note that correlations involving WERs did not reach statistical significance. With a larger compute budget, it might be possible to resolve the variance in WERs. In the present work, we used about 1,000 GPU hours (279 training LMs and 730 evaluating Whisper models).

Table 1: **Example ASR predictions and ground truth sentences suggest directions for future work.** Here we show three predictions from the best Whisper model (`large-v2`) rescored with the best LM (1/1 training data). To facilitate the discussion in the text, character-level edits are colored for deletion, substitution, and insertion. Differences in capitalization and punctuation are ignored, as the text was normalized before evaluation. Elsewhere in the paper, WERs were computed at the word-level. See text for discussion.

ID	ASR Prediction	Ground Truth
1	O Lahaina ke kapikala hiko o Hawai‘i.	‘O Lāhaina ke kapikala kahiko o Hawai‘i.
2	Makemake wau e ike i nā wai e hā.	Makemake wau e ‘ike i Nā Wai ‘Ehā.
3	Ai, e hele māka i ka ino kaua.	‘Ae. E hele māka‘ika‘i nō kāua.

4 Discussion

There are dozens of hours of labeled Hawaiian audio but millions of pages of Hawaiian newspaper text available right now. We therefore investigated, in this paper, how the use of text data might improve Hawaiian ASR. We also evaluated the capacity for a foundation model, Whisper (2), to accurately transcribe Hawaiian audio without having seen any labeled Hawaiian data. Out of the box, we found the largest Whisper models (`large` and `large-v2`) could transcribe Hawaiian audio with WERs of about 22% (Figure 1). By incorporating a Hawaiian LM, the `large-v2` model achieves a WER of about 20% (Figure 2). Ultimately, the difference in WER is only about 1–2% but the improvement is consistent across repeated evaluations ($p < 0.01$). For low-resource languages like Hawaiian, which could use better ASR to accelerate language preservation and revitalization efforts, any statistically significant improvement in performance is welcome. But there is more work to do.

To suggest ways of improving on these results, let us consider where they fail. In Table 1, we see three ASR predictions from the Whisper `large-v2` model together with corresponding ground truth transcriptions. Character-level errors have been highlighted to indicate deletions, substitutions, and insertions when mapping from the prediction to the ground truth transcript. A number of intriguing patterns emerge. For example, we see that the model frequently fails to capture phonemic glottal stops /P/ (written in Hawaiian with the ⟨‘⟩ symbol, which in Hawaiian is called an *‘okina* /Pokina/). The phoneme /P/ can be realized in Hawaiian speech with full closure, modal voice, or creak (15; 16; 17). Acoustically, these are relatively quiet sounds. In a few cases (the glottal stops in *māka‘ika‘i*), we see that the model tries to capture these with spaces. This may relate to a bias for English phonology to insert a glottal stop before a word-initial vowel. This would also account for the missing glottal stop on *‘Ae*. That said, not all of the extra spaces added by the model (e.g. *e hā*, *i ka*) correspond to missing glottal stops and perhaps show that the model has not properly learned to segment Hawaiian words. Or, since *e*, *hā*, *i*, and *ka*, are all attested Hawaiian words (18), it may be more accurate to say the model has not learned to predict the correct word given the context.

Another interesting failure mode is the treatment of vowels. In many cases, the model fails to distinguish phonemically short and long vowels (e.g. *Lāhaina* /la:.hai.na/, *nō* /no:/, *kāua* /ka:u.a/).[2] Noticeably, the model also gets the vowel wrong on the diphthong *‘Ae*. We note that the quality of a vowel in Hawaiian can differ whether in a monophthong or diphthong (15; 19). All of these errors suggest that the model does poorly on sounds that diverge from English, which makes sense given that Whisper is optimized for English. We hypothesize that these errors may be corrected by fine-tuning Whisper on sufficient amounts of labeled Hawaiian data. Serious efforts are currently underway within the Hawaiian community to gather and organize a much larger collection of labeled Hawaiian data and we are actively working to leverage all of it to improve Hawaiian ASR.

Following the principle of using all the data at hand, the Hawaiian community has access to a lot more unlabeled text, and unlabeled audio. On the text side, one might train a better LM, or even LLM, for Hawaiian (e.g. using Transformers (20)) while scaling up the size of the training set. One might also advance text normalization methods (6; 7). These could then be used to modernize a larger amount of the 19th century texts that were written in an older missionary orthography which excludes letters for Hawaiian sounds that do not occur in English (21). Finally, two obvious ways to incorporate unlabeled audio would be to use self-supervised learning (22; 23; 24; 25; 3; 26; 27; 28) and pseudo-labeling (29; 30; 31; 32; 33; 34).

What can other low-resource language communities learn from this study? We conjecture that many languages are like Hawaiian in having much less labeled data than unlabeled data (e.g. text corpora).

[2] *Lāhaina* is also written and pronounced without the long vowel: *Lahaina*.

Part of our motivation for including the posthoc experiments, which varied the amount of text data used to train an LM, was to provide a guide for languages with less text data. Nonetheless, when it comes to leveraging models like Whisper, Hawaiian may not be representative. For example, other languages may not be included in Whisper's training data. After all, Whisper was trained on 338 hours of unlabeled audio from Hawaiian as well as 1381 hours of unlabeled audio from Māori, a closely related Polynesian language(2). The Hawaiian alphabet is also very similar to English, with only a few modifications for glottal stops and long vowels. Ultimately, how well our approach works for other low-resource languages is an empirical question. But that may be enough. For languages that have few options, we offer hope: it is worth trying, as Hawaiian tradition teaches, to use all the data you have.

5 Limitations

The present work is limited in its choice of language model (LM) and foundation model, and in its ability to make strong claims about other low-resource languages. We chose to use an LSTM-based LM to replicate prior work (6; 7), which allowed us to avoid optimizing the model architecture. However, as Transformers dominate the current LM landscape, we believe future work will benefit by exploring alternative LM architectures (20). Second, using Whisper as our foundation model limits the granularity with which we can interpret results. For example, we do not know what Hawaiian (or Māori) data OpenAI used to train Whisper, and therefore can only hypothesize broadly about how this data affected the model performance. A significant concern is the potential overlap between Whisper's training data and our ASR Test Set. However, we believe this overlap is unlikely. Hawaiian was used in Whisper's translation task (2), mapping Hawaiian audio to English transcripts, and to our knowledge there are no English transcripts of the audio recordings we used. However, we strongly encourage OpenAI to release more information about their training data, for the benefit of humanity. Finally, while our post-hoc analyses were aimed at providing insights for speakers of other low-resource languages, it remains an empirical question how effective our approach of rescoring a foundation model like Whisper with a custom LM will be for other langauges. We look forward to seeing this approach applied to other languages in future research.

References

[1] M. K. Pukui, *ʻŌlelo Noʻeau: Hawaiian Proverbs & Poetical Sayings*. Honolulu: Bishop Museum Press, 1983.

[2] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *arXiv preprint arXiv:2212.04356*, 2022.

[3] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "Wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in Neural Information Processing Systems*, pp. 12 449–12 460, 2020.

[4] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, "XLS-R: Self-supervised cross-lingual speech representation learning at scale," *INTERSPEECH*, pp. 2278–2282, 2022.

[5] G. Sun, X. Zheng, C. Zhang, and P. C. Woodland, "Can contextual biasing remain effective with Whisper and GPT-2?" *INTERSPEECH*, 2023.

[6] B. Shillingford and O. Parker Jones, "Recovering missing characters in old Hawaiian writing," in *Empirical Methods in Natural Language Processing (EMNLP)*, 2018, pp. 4929–4934.

[7] O. Parker Jones and B. Shillingford, "Composing RNNs and FSTs for small data: Recovering missing characters in old Hawaiian text," *NeurIPS Workshop on Interpretability and Robustness in Audio, Speech, and Language (IRASL)*, 2018.

[8] D. Jurafsky and J. Martin, "Speech and language processing," 2024, Feb 3, 2024 draft. [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/

[9] Hoʻoulumāhiehie, *Ka Moʻolelo o Hiʻiakaikapoliopele*, M. P. Nogelmeier, Ed. Honolulu: Awaiaulu, 2021, originally published between 1905 and 1906 as a daily series in the Hawaiian-language newspaper *Ka Na'i Aupuni*.

[10] I. Sutskever, J. Martens, and G. Hinton, "Generating text with recurrent neural networks," *International Conference on Machine Learning (ICML)*, pp. 1017–1024, 2011.

[11] A. Graves, "Generating sequences with recurrent neural networks," *arXiv preprint arXiv:1308.0850*, 2013.

[12] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[13] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[14] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *International Conference on Learning Representations (ICLR)*, 2015.

[15] O. Parker Jones, "Hawaiian," *Journal of the International Phonetic Association*, vol. 48, no. 1, pp. 103–115, 2018.

[16] L. Davidson, "Effects of word position and flanking vowel on the implementation of glottal stop: Evidence from Hawaiian," *Journal of Phonetics*, vol. 88, p. 101075, 2021.

[17] L. Davidson and O. Parker Jones, "Word-level prosodic and metrical influences on Hawaiian glottal stop realization," *Phonetica*, vol. 80, no. 3–4, pp. 225–258, 2023.

[18] M. K. Pukui and S. H. Elbert, *Hawaiian Dictionary*. Honolulu: University of Hawai'i Press, 1986, Revised and Enlarged Edition.

[19] T. Kettig, "Ha'ina 'ia mai ana ka puana: The vowels of 'ōlelo Hawai'i," Ph.D. dissertation, University of Hawai'i at Mānoa, 2021.

[20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Łukasz Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[21] A. J. Schütz, *The Voices of Eden: A History of Hawaiian Language Studies*. Honolulu: University of Hawai'i Press, 1994.

[22] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[23] S. Schneider, A. Baevski, R. Collobert, and M. Auli, "wav2vec: Unsupervised pre-training for speech recognition," *INTERSPEECH*, pp. 3465–3469, 2019.

[24] Y.-A. Chung, W.-N. Hsu, H. Tang, and J. Glass, "An unsupervised autoregressive model for speech representation learning," *INTERSPEECH*, pp. 146–150, 2019.

[25] A. Baevski, S. Schneider, and M. Auli, "vq-wav2vec: Self-supervised learning of discrete speech representations," *International Conference on Learning Representations (ICLR)*, 2020.

[26] Y.-A. Chung, Y. Zhang, W. Han, C.-C. Chiu, J. Qin, R. Pang, and Y. Wu, "W2v-BERT: Combining contrastive learning and masked language modeling for self-supervised speech pre-training," *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 244–250, 2021.

[27] W.-N. Hsu, Y.-H. H. Tsai, B. Bolte, R. Salakhutdinov, and A. Mohamed, "HuBERT: How much can a bad teacher benefit ASR pre-training?" *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6533–6537, 2021.

[28] W.-N. Hsu, B. Bolte, Y.-H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[29] D.-H. Lee, "Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks," *ICML Workshop on challenges in representation learning*, vol. 3, 2013.

[30] G. Synnaeve, Q. Xu, J. Kahn, T. Likhomanenko, E. Grave, V. Pratap, A. Sriram, V. Liptchinsky, and R. Collobert, "End-to-end ASR: from supervised to semi-supervised learning with modern architectures," *ICML Workshop on Self-supervision in Audio and Speech*, 2020.

[31] D. S. Park, Y. Zhang, Y. Jia, W. Han, C.-C. Chiu, B. Li, Y. Wu, and Q. V. Le, "Improved noisy student training for automatic speech recognition," *INTERSPEECH*, pp. 2817–2821, 2020.

[32] Q. Xu, T. Likhomanenko, J. Kahn, A. Hannun, G. Synnaeve, and R. Collobert, "Iterative pseudo-labeling for speech recognition," *INTERSPEECH*, pp. 1006–1010, 2020.

[33] T. Likhomanenko, Q. Xu, J. Kahn, G. Synnaeve, and R. Collobert, "SlimIPL: Language-model-free iterative pseudo-labeling," *INTERSPEECH*, 2021.

[34] D. Berrebbi, R. Collobert, S. Bengio, N. Jaitly, and T. Likhomanenko, "Continuous pseudo-labeling from the start," *International Conference on Learning Representations (ICLR)*, 2023.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction summarize the paper's contribution and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: A limitations section is included.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA] .

   Justification: The paper produces no theoretical results.

   Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The methods section discloses the steps taken to produce the paper's results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The purpose of the paper was to explore methods for improving ASR for Hawaiian using text-only data rather than share the explicit language model and data that were used.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The methods section details the training details and hyperparameters chosen.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper reports error bars.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.

- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The paper reports total GPU hours used

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Authors have read the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The introduction and discussion sections both discuss the significance that this work holds for low-resource language ASR.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA] .

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Permission was obtained to use the datasets used in this work, and the models used are properly cited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [No]

    Justification: The rescoring method used in the paper is an established technique. The main contribution of the paper is the results of how Whisper performs on Hawaiian after rescoring with a Hawaiian language model.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA] .

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA] .

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.