Solving Failure Modes in the Creation of Trustworthy Language Technologies

Gianna Leoni, Lee Steven, Miles Thompson, Tūreiti Keith, Keoni Mahelona, Peter-Lucas Jones, Suzanne Duncan
Te Reo Irirangi o Te Hiku o Te Ika (Te Hiku Media)
1 Melba Street, Kaitāia, Aotearoa New Zealand
{gianna, lee, miles, tureiti, keoni, peterlucas, suzanne}@tehiku..co.nz

Abstract
To produce high-quality Natural Language Processing (NLP) technologies for low-resource languages, authentic leadership and participation from the low-resource language community is crucial. This reduces chances of bias, surveillance and the inclusion of inaccurate data that can negatively impact output in language technologies. It also ensures that decision-making throughout the pipeline of work centres on the language community rather than only prioritising metrics. The NLP building process involves a range of steps and decisions to ensure the production of successful models and outputs. Rarely does a model perform as expected or desired the first time it is deployed for testing, resulting in the need for re-assessment and re-deployment. This paper discusses the process involved in solving failure modes for a Māori language automatic speech recognition (ASR) model. It explains how the data is curated and how language and data specialists offer unparalleled insight into the debugging process because of their knowledge of the data. This expertise has a significant influence on decision-making to ensure the entire pipeline is embedded in ethical practice and the work is culturally appropriate for the Māori language community thus creating trustworthy language technology.

Keywords: Māori language, language technologies, ethics, automatic speech recognition

1. Introduction

Te reo Māori (the Māori language) has deep Polynesian roots as an oral language. It was the language of wider communication in New Zealand up until its rapid decline between 1900 and 1950 (Leoni, 2016). It is beyond the scope of this paper to go in-depth regarding the loss of the language. However, it is important to note that there have been deliberate attempts to colonise and assimilate Māori and remove the Indigenous language from its people for over 150 years (see Higgins et al., 2014; Keenan, 2012; Winitana, 2011; Walker, 1990; Te Rito, 2008). Despite significant progress since the 1970s to revitalise the language, and work that is often replicated by other Indigenous peoples, there are still many issues relating to Māori language communication. Speaker numbers are low, there is a lack of adequate resources available, and there is limited high-quality technology that allows for Māori language engagement. This impacts the writing, speaking, listening and reading of the language with everyday devices that are meant to make peoples' lives easier (Te Reo Irirangi o Te Hiku o te Ika, 2022).

Natural language processing (NLP) enables computers to understand human speech, but how it functions positively for high-resource languages is very different to low-resource languages (Barss, 2019). For te reo Māori, much of this relates to the absence of high-quality large Māori language data sets that are needed for machine learning (ML). Te reo Māori was only written for the first time in the early 1800s and the continual attempts to eradicate the language in favour of English has impacted language user capacity. This has resulted in limited sources of language data in te reo Māori compared to high-resource languages.

It is hard for organisations dedicated to low-resource languages like te reo Māori to compete with Big Tech in the pursuit of quality language technology tools. These Big Tech companies have the people, money, and data (often unethically sourced). Many also lack appropriate standards that facilitate the creation of ethically appropriate NLP tools. Furthermore, natural language processing tools are rarely developed by Indigenous peoples with an Indigenous perspective. This method leads to poor-quality outputs that often cause harm to low resource languages as there is a lack of transparency in the process, they often breach privacy standards or surveil people, they are full inaccuracies, and they perpetuate negative biases (Jones et al., 2023; Dubay & Nalbandian, 2021).

Te Reo Irirangi o Te Hiku o te Ika (Te Hiku Media), a tribal radio station based in Kaitāia, New Zealand, has been on a mission to create ethically sound and culturally appropriate tools for indigenous languages (Te Reo Irirangi o Te Hiku o te Ika, 2022). This starts with how data is collected and curated, to how it is used in data processing, engineering, addressing failure modes, finetuning and output.

A significant part of the journey has been recognising that using high quality data is paramount to Te Hiku Media's success. This is further supported by having a thorough process where dialogue occurs between the data and language specialists and the data scientists and ML engineers (Jones et al., 2023b). This is particularly useful when it comes to debugging failure modes.

This paper discusses the process involved in solving failure modes for Te Hiku Media's Māori language speech-to-text (STT) model. It highlights how language and data specialists offering insight into the problem-solving process strengthen the cultural integrity of the model. It first explains the significance of knowing the data and the curation process. This is followed by a brief description of how the data was used for this particular project. Finally, the paper outlines how Te Hiku Media debugs failure modes and the discussion and decision-making that occurs. This part of the process consolidates the ethical

practice and ensures the work is culturally appropriate for the language community it is being made for.

2. The Data

Te Reo Irirangi o Te Hiku o te Ika has been collecting and archiving content from its broadcasting activities since 1990 (Te Reo Irirangi o Te Hiku o te Ika, 2022). Whilst maintaining its radio presence Te Hiku Media has expanded to include online TV and data science technology development, all of which are committed to the revitalisation of the Māori language. This audio and audio-visual content now forms the basis of the largest archive in the tribal radio network. The protection of the knowledge and content in the archive was intentional (Jones et. al., 2023a) and abiding by cultural protocols to ensure it was cared for was natural.

Te Hiku Media has been developing innovative and Indigenous-led solutions to enable Indigenous peoples to engage with the digital world while also protecting Indigenous knowledge and ensuring data sovereignty. All of the work is guided by the communities where Te Hiku Media is based. This ensures that it is ethically and culturally appropriate regardless of the ethnicity of any practitioners working on any projects. It is important to note, however, that Te Hiku Media prioritises the hiring of Indigenous staff. Two (out of three) of the Executive team are Māori and are genealogically linked to the region of Te Hiku o Te Ika, and the other is Hawaiian. Furthermore, 80% (9 out of 11) of those working on the data science project are Indigenous.

The content collected over the past 33 years provides a unique source of knowledge and data that can be used for Te Hiku Media's data science endeavours. Whilst a large data source in terms of anything similar available in te reo Māori, it is much smaller than what is usually required for NLP, ML and automatic speech recognition (ASR). Now that the data has been digitised and made accessible, it has become increasingly obvious people within the team must have an intimate knowledge of the data and using the data in culturally appropriate ways will positively impact any output.

2.1 Knowing the Data

Jones et al. (2023b) discuss how Te Hiku Media's prioritisation of Māori language expertise has been a key factor in the success of the work programme and contributes to maintaining ethical space for Indigenous peoples. Data and language specialists are responsible for transcribing, reviewing and confirming the suitability of audio content before it can enter a training, test or validation dataset, a task known as labelling data. This is often a time-consuming task, that contrasts much of what is expected in the world of NLP where technology is being built to save time. However, carefully curating the datasets ensures that the data input is high quality and intelligible which could negatively impact any output. This has a flow-on effect on the curating of datasets for particular projects. If the ultimate aim is to exemplify a native speaker sound, with the type of language and prosody that would be viewed as aspirational for second language learners today, the Māori language specialists can advise which data to use from the archive. If a project aims to transcribe a range of voices, the team will ensure a fair representation of gender and age and native and second-language speakers from different tribes in New Zealand (Jones et al., 2023b).

An ongoing issue for under-resourced languages is the lack of quality data available to create tools. Many attempts (especially by Big Tech) create bias in the language outputs (Dubay & Nalbandian, 2021). This usually occurs in large, uncurated and/or unethical datasets. For example, poor training data might reiterate grammatical errors; biased training data may reinforce negative and harmful stereotypes about indigenous or minority groups; and unethical training data has likely been taken or used without permission. However, when care is taken throughout the data curation process this ensures that data is respected and Indigenous knowledge is protected. Offensive or unsuitable data can be removed before training occurs, limiting opportunities for bias or offending people. Indigenous knowledge that is not open information can be preserved and only shared with those who should have it (Jones et al., 2023b).

2.2 Using the Data

Jones et al. (2023b) introduce Te Hiku Media's pipeline in developing an ASR model. The ultimate aim of the ASR model is to contribute to the restoration of the Māori language by exemplifying a native speaker sound, that is, the type of language and prosody that would be viewed as aspirational for second language learners today.

Of particular importance in the ASR model work is Te Hiku Media's STT model. Initially created for te reo Māori, it has been through several iterations since its creation. Originally built using Mozilla's Deepspeech architecture (Hannun et al., 2014), which relied on recurrent neural networks (RNN), the model has since transitioned to Nvidia's implementation of Conformer, a convolution-augmented transformer (Gulati et al., 2020). Moreover, the evolution of the STT model is not limited to architectural enhancements alone; there has been a substantial expansion in the corpus of training data utilised, growing from approximately 400 hours to 5000 hours.

Alongside the architectural improvements and data augmentation, the performance metrics of the STT model have demonstrated significant progress. The word error rate (WER) is measured against the custom-curated dataset of labelled target sentences specifically designed for benchmarking automatic speech recognition performance on te reo Māori (Jones et al., 2023). These are quantitatively analysed to see if the proposed model is better, the same or worse, and how accurate it is. There has been a substantial drop in the WER of the STT model from 27% to 10%.

Once a WER report is created, the language and data specialists also analyse the target sentences qualitatively, as the WER is not always an accurate

indicator because of language nuances. Sometimes a WER report will suggest that the proposed model is performing better or worse, but the language and data specialists can see that linguistically the model is producing the opposite (see Jones 2023b for an example). Accepting the WER in these instances without careful consideration could negatively impact the language community.

Once the 3-4 data scientists and ML engineers have been able to quantitatively review the WER report and the 2-3 data and language specialists have qualitatively reviewed the WER report a meeting is set up to discuss the findings. This collaborative approach is often how failure modes are discovered.

3. Addressing Failure Modes

Failure modes in NLP refer to the many ways models can fail to perform as expected or desired this could be either technical or linguistic. In the ASR development, failure modes show when the model has poor performance in language domains, this includes grammar, regional variations in language, or different types of speech (like songs, a radio interview or a formal speech). Whilst it can be reassuring that some language domains work well, a model that fails to transcribe an important element of the Māori world correctly is not ethically or culturally appropriate and impacts the overall quality of the model despite the WER.

Addressing these failure modes requires a combination of work from data scientists, ML engineers and data and language specialists. Which failure modes should receive attention is discussed at the collaborative meetings and the particular process required moving forward is decided on. It usually includes examining the data processing, model selection, finetuning, monitoring and maintaining already deployed models. It also requires analysis of the data that has been curated. Because Te Hiku Media has a comprehensive understanding of the data, it is better prepared to debug failure modes.

A significant failure mode in the current STT model has been the loss of text when speaker-switching occurs in an interview or conversation. The transcript struggles to pick up the second speaker's voice, but if the first speaker returns, it can recognise it and will continue transcribing.

3.1 The Process

To address the initial failure mode, the data specialists created a 60-minute dataset that was specifically designed to provide examples of speaker switching.

Ngā Take o Te Taitokerau is a radio news segment from Te Reo Irirangi o Te Hiku o te Ika (Te Hiku Media, n.d.). The segments are approximately 4 minutes long and usually include a presenter, as well as two voice clips from interviews that were conducted on air during that day. The language specialists agreed that this would provide a dataset with ample examples of speaker-switching to gauge how the current and any proposed models functioned.

The segments were uploaded to Kaituhi (Te Hiku Media's transcription platform) and then split/unsplit accordingly, and transcribed once in their respective splits. They were presented in four different formats.

1) Full 4-minute segment, edited verbatim
2) Full 4-minute segment, model transcription
3) 4x <60second segments, edited verbatim
4) 4x <60second segments, model transcription

The full segments showed how the model would cope with a longer piece of audio and the <60second segments demonstrated the models handling of shorter, more concise audio inputs. The 60-second segments were chosen as this is the maximum duration for training data, ensuring consistency and relevance to the model's capabilities. The edited segments were reviewed by the language specialists to ensure that the Māori language was correct and that there would be an accurate view of any disparities. Throughout the curation process, it became clear that this was indeed a verified method to test this failure mode.

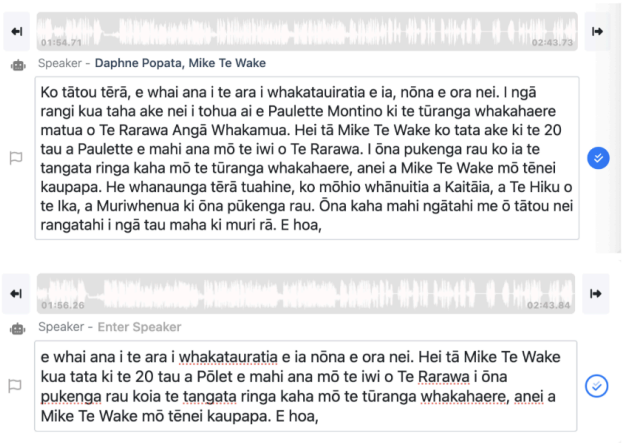


Figure 1: Example of <60s segment with a loss of text when speaker switching occurs

In addressing the identified failure mode, the importance of curating additional training data that specifically showcased speaker switching scenarios became increasingly apparent. To achieve this, data augmentation techniques were employed, leveraging the concatenation of audio and transcripts sourced from different speakers. This approach aimed to enrich the training dataset with diverse examples of speaker transitions, thereby enhancing the model's ability to accurately transcribe such instances, and contributing to its overall robustness.

An initial issue identified by the language specialists was that the model struggled with longer numbers and writing these as numerals (the chosen orthographic convention of the project for numbers). This appeared in the data curation process when the model was unable to complete the transcription. The model's difficulty with accurately transcribing long numbers may stem from a combination of factors.

Firstly, large numbers are infrequently represented in the training dataset, limiting the model's exposure to proper numeral formats. Inconsistent formatting within the training data, such as the insertion of spaces after commas (e.g 1, 000, 000) in large numbers, contrasts with the standard numeral format, potentially confusing the model. Additionally, the use of Byte Pair Encoding (BPE) tokenization could exacerbate this issue by segmenting these inconsistently formatted numbers in unpredictable ways.

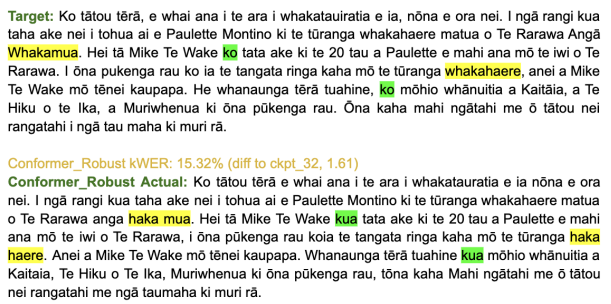
3.2 Analysis

The curation and re-testing then required analysis and discussion. This included both qualitative and quantitative WER analyses.

For Te Hiku Media, this is when the team of data specialists (2-3 people), data scientists and ML engineers (3-4 people) once again review reports and then meet to discuss the information presented. All members need to be included as everyone provides different knowledge and views of what might be causing the failure modes and if there are any other perceived problems in the model. Thorough discussion improves the whole team's understanding of the pipeline of work. This allows team members to have a better awareness of what they might need to do when making adjustments to their particular area of work. It also makes decision-making more effective because fewer assumptions will be made about the different parts of the work.

Despite mainly addressing the issue of the missing language when speakers switch and improving the WER as a result, the new model produced several more failure modes.

Upon thorough analysis of the report, the language specialists realised that the model had broken recognition of a significant aspect of the language variation of Te Taitokerau. In Te Taitokerau, it is common and natural for native speakers to pronounce the digraph 'wh' as 'h'. For example, 'whakarongo' (to listen) is pronounced 'hakarongo'. In the International Phonetic Alphabet, this distinction can be represented as the 'wh' [f] sound being pronounced as [h]. However, orthographic conventions dictate that the word is still spelt 'whakarongo' despite whether it is pronounced 'fakarongo' or 'hakarongo'. This original spelling and voice recognition of the language variation of Te Taitokerau had never been an issue in previous models. A voice would say 'hakarongo' and the ASR would produce the word in written form as 'whakarongo'. However, in the most recent report, the model started producing this inconsistently. For example, it removed the 'w' and split up the word, e.g. the verb 'whakamua' (forward, ahead) became 'haka mua' (haka = te perform/dance, mua = forward; no linguistically logical translation apparent) (see yellow highlighting in Figure 2). In essence, it's produced gibberish.


Target: Ko tātou tērā, e whai ana i te ara i whakatauiratia e ia, nōna e ora nei. I ngā rangi kua taha ake nei i tohua ai e Paulette Montino ki te tūranga whakahaere matua o Te Rarawa Angā Whakamua. Hei tā Mike Te Wake ko tata ake ki te 20 tau a Paulette e mahi ana mō te iwi o Te Rarawa. I ōna pukenga rau ko ia te tangata ringa kaha mō te tūranga whakahaere, anei a Mike Te Wake mō tēnei kaupapa. He whanaunga tērā tuahine, ko mōhio whānuitia a Kaitāia, a Te Hiku o te Ika, a Muriwhenua ki ōna pūkenga rau. Ōna kaha mahi ngātahi me ō tātou nei rangatahi i ngā tau maha ki muri rā.

Conformer_Robust kWER: 15.32% (diff to ckpt_32, 1.61)
Conformer_Robust Actual: Ko tātou tērā e whai ana i te ara i whakatauratia e ia nōna e ora nei. I ngā rangi kua taha ake nei i tohua ai e Paulette Montino ki te tūranga whakahaere matua o Te Rarawa anga haka mua. Hei tā Mike Te Wake kua tata ake ki te 20 tau a Paulette e mahi ana mō te iwi o Te Rarawa, i ōna pūkenga rau koia te tangata ringa kaha mō te tūranga haka haere. Anei a Mike Te Wake mō tēnei kaupapa. Whanaunga tērā tuahine kua mōhio whānuitia a Kaitaia, Te Hiku o Te Ika, Muriwhenua ki ōna pūkenga rau, tōna kaha Mahi ngātahi me ō tātou nei rangatahi me ngā taumaha ki muri rā.

Figure 2: Example of 'h' vs 'wh' (yellow) and 'ko' vs 'kua' (green) in report

Another output that had not previously been an issue was the model's recognition of 'ko' and 'kua' and mixing these up (see green highlighting in Figure 2). These words are used as tense markers at the beginning of a sentence and are usually followed by a verb (ko can be followed by many things). Previously the model had been far more accurate in determining the difference.

3.3 The Decision-Making

Whilst word error rates may be high, certain necessary and sufficient conditions must be adhered to when considering sending models to production. As an Indigenous-led data science team, Te Hiku Media have always prioritised authentic and high-quality model outputs, and this ultimately influences all final decisions. This links back to how Te Hiku Media is guided by its community, therefore if the data and language specialists do not believe that a model is ready, more work must be done before it is sent to production.

The two orthographic failure modes in Section 3.2 provide useful examples of what is a necessary condition and what is a sufficient condition. The 'ko' vs 'kua' issue has emerged, but grammatically the output could be either word and still be correct. Upon listening to the audio with the target and suggested transcriptions, this would be a sufficient condition. Whilst this failure mode will be worked on for future iterations, it would not hold up deploying the model to production. This is because the result is not ungrammatical or produces an issue that might cause a negative reaction from the language community.

The 'wh' [f] vs 'h' [h] no longer working is a more serious issue. As Te Hiku Media is guided by and responsible to the five tribes of the Far North of New Zealand, this function failing does not accurately reflect the language community this model is being built for. In previous iterations, the model was capable of processing words said like 'hakarongo' and spelling them as 'whakarongo' [fakarongo]. This becomes a non-negotiable and necessary condition because WER decrease is not more important than

the language community and how they are represented. It meant that further work needed to be completed before the model could be pushed to production.

3.4 The Next Steps

After the decision was made that further work was needed to rectify the [f] and [h] issues, the team focused more attention on the data curation choices in an attempt to debug and resolve this issue. The team speculated that the extra 930 hours of English audio/text pairs added to the training dataset may have contributed to this problem. Bilingual performance across both English and Māori is an important goal for the ASR, but the expanded English data potentially included more sound/text pairs where the [h] sound maps to the token 'h' ('English' words like 'happy', 'hazel' or 'haute' for example). This stronger English association may have loosened the association for regional variants where an audible pronunciation of the [h] needs to map to 'whakarongo'.

The failure mode was resolved by balancing this out with an additional ~90 hours from the Kōrero Māori project, which includes a lot of regional variation contributed from around Aotearoa (Te Reo Irirangi o Te Hiku o te Ika, 2022), as well as adding 500+ hours of synthetic code switching data. Another positive result was the further improvement of WER on the benchmarks during latest testing and release.

By reflecting and collaborating as a whole team, the failure mode was rectified to the point where the model could be pushed to production because all members of the team (and in particular the language and data specialists) were satisfied with the latest WER report results. The process undertaken to reach this point emphasises the importance of having qualitative evaluation, sensitive to important issues such as performance under regional variation, included as part of the core of the work in iterating on and improving the language model. As a result, in improved the overall WER whilst maintaining quality and ethnically appropriate language outputs.

4. Conclusion

It has become increasingly important to Te Hiku Media to create trustworthy, authentic, dependable and ethical tools. First and foremost, this is to ensure that the language community the tech is being created for is represented, both in creating the tech and in the output and usability of the tech. The threat to high-quality language technology for under-resourced languages is growing.

The attention given to the discussion and decision-making when addressing failure modes ensures the building of quality products that are culturally appropriate for under-resourced language communities like te reo Māori. The process undertaken by Te Hiku Media guarantees that it positively contributes to te reo Māori revitalisation rather than causing harm or reinforcing grammatical errors. Te Hiku Media will not blindly follow metrics or good WER if it is detrimental to the overall quality of the language output or will negatively impact the language community.

5. Acknowledgements

We acknowledge the continuing support from the five tribes, the trustees that represent them and the many community members of Te Hiku o te Ika that contribute to our work.

This work was funded by the New Zealand Ministry for Business, Innovation and Employment through the Strategic Science Investment Fund.

6. Bibliographic References

Barss, P. (2019) Can we eliminate bias in AI? How Canada's commitment to multiculturalism could help it become a world leader, U. of T. News. https://www.utoronto.ca/news/can-we-eliminatebias-ai-how-canada-s-commitment-multiculturalism-could-help-it-become-world, accessed on 14 October 2022

Dubay, L. and Nalbandian, L. (2021). Creating an equitable AI policy for Indigenous communities, First Policy Response. https://policyresponse.ca/creating-anequitable-ai-policy-for-indigenous-communities/, accessed on 14 October 2022

Gulati, A., Qin, J., Chiu, C., Parmar, N., Zhang, Y., Yu, J., Han, W., Wang, S., Zhang, Z., Wu, Y., & Pang, R. (2020). Conformer: Convolution-augmented Transformer for Speech Recognition. *https://arxiv.org/abs/2005.08100*

Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., & Ng, A. Y. (2014). Deep Speech: Scaling up end-to-end speech recognition. *https://arxiv.org/abs/1412.5567*

Higgins, R, Rewi, P., and Olsen-Reeder, V. (Eds.). (2014). *The value of the Māori language – te hua o te reo Māori*, Wellington: Huia Publishers.

Jones, P-L., Mahelona, K., Duncan, S., and Leoni, G. (2023a). Ngā taonga tuku iho: Intergenerational transmission using archives. *Ethical Space: International Journal of Communication Ethics*, 20(2/3).

Jones, P.-L., Mahelona, K., Duncan, S., and Leoni, G. (2023b). Kia tangata whenua: Artificial intelligence that grows from the land and people. *Ethical Space: International Journal of Communication Ethics*, 20(2/3).

Keenan, D. (2012). *Huia histories of Māori: ngā tāhuhu kōrero*, Wellington, Huia Publishers.

Leoni, G. (2016). *Mā te taki te kāhui ka tau*. Unpublished PhD thesis, Dunedin, University of Otago.

Te Hiku Media. (n.d.) Ngā Take o Te Taitokerau. *Te Hiku Media.* https://tehiku.nz/te-reo/nga-take/

Te Reo Irirangi o Te Hiku o te Ika. (2022). He Reo Tuku Iho, He Reo Ora: Living language transmitted intergenerationally. *Mai Journal*, 11(1).

Te Rito, Joseph (2008) Struggles for the Māori language: He whawhai mo te reo Māori, *MAI Review*, 2:1-8.

http://www.review.mai.ac.nz/mrindex/MR/article/view/164.html.
Walker, R. (1990). *Ka whawhai tonu mātou – struggle without end*, Auckland: Penguin Books.
Winitana, C. (2011). *Tōku reo, tōku ohooho: Ka whawhai tonu mātou*, Wellington: Huia Publishers.

7. Language Resource References

Moorfield, J.C. 2005. *Te Aka: Māori-English, English-Māori Dictionary and Index*. Auckland: Pearson-Longman.