NLP Technologies to Support Learning in Low Resource Languages

Cat Kutay[1], Tamati Smith[2], Nabin Khatawada[1], Sabin Ghimire[1],
[1]Charles Darwin University, Darwin, Australia, [2]Utakarra WA, 6530
Correspondence: cat.kutay@cdu.edu.au

Abstract

Australian Aboriginal and Torres Strait Islander languages are the oldest surviving languages in the world. They teach intricate knowledge about the environment, technology, astronomy, health, and governance. Despite many obstacles these oral forms continue to be spoken and are being reclaimed by those who were forbidden to speak their languages for years. This paper presents some of the technology being developed with communities to support oral language use and text translations. We follow ICIP principles in demonstrating the use of Augmented Intelligence (AI) approaches that suit their local culture and history. We describe the outcomes of these first AI models in terms of material used and the community perceived value despite some concerns that have been raised about their suitability relating to the use of recognised voices, limited grammatical modelling and lack of sufficient context of text based material. We plan to continue this work on open-source software pipelines facilitate language learning, documentation, and language practice for learners. Increasing the respect for and usage of these languages online through community control may also strengthen their use in younger generations within their community.

Keywords: speech recognition, text to speech, AI translations

1 Introduction

1.1 History of support

Global recognition of the importance of the preservation of Indigenous languages as living resources of knowledge, perspectives and values has increased the focus on computing support that does not remove Data Sovereignty (ID-Sov) and Indigenous Cultural and Intellectual Property rights (ICIP). Also as many Indigenous languages are used in remote areas that retain high biodiversity and/or retain knowledge of local plants and environmental managements (Romaine and Nettle, 2000), the economic benefit of the strengthening of these languages is recognised (Hamans et al., 2024). However the benefits to the language diaspora is much broader and encompasses all aspects of well-being (Dianati and Bolt, 2025).

This work was developed through a network of First Nations working with Augmented Intelligence (AI) and languages to support: First Nations studying at university to engage with learning material translated into their oral language; those studying their local languages to hear and practice their speaking; text interpretation for language speakers and written language tutors which translates with comments and examples.

1.2 Language analysis

For many languages the resources we have to reclaim or revitalise languages are text material collected by missionaries, soldiers and settlers who had a profound interest in the ways of thinking and speaking by their neighbours in early Australian communities. While often the words were not translated with the same significance held by the original speakers, there was some attempt to gather words. However, wordlists pulled the data out of context and we are finding that Australian Aboriginal languages are highly contextualised, in that words will have many meanings depending on the country where they are spoken, who is addressed, the story or topic of conversation.

Language restoration led by First Nations communities aims to restore or strengthen their languages. Through many initiatives, including master-apprentice classes (Commonwealth of Australia, 2020), First Nations communities work to recover lost linguistic practices, promote intergenerational transmission, and ensure that these languages remain living systems. Furthermore,

teaching non-First Nations service providers how to communicate in the local language aims to improve the service they will provide.

Despite the potential of AI in language revitalisation, the construction of AI systems for endangered languages presents significant challenges. Many First Nations languages, including Australian Aboriginal and Torres Strait Islander languages, have limited available data in the form of recorded speech or written texts, which hinders the development of robust AI models. In addition, these languages often lack standardised orthographies, making it difficult to create consistent training data. Cultural and ethical considerations also play a critical role, so co-design with First Nations communities must be maintained to ensure that language knowledge is handled with respect and aligns with community goals (Australian Research Data Commons (ARDC), 2024).

1.3 Research Outline

The aims of this research are to investigate the value or not of the new large language models (LLMs) in the context of community based language reclamation and strengthening. While these models are trained on many human languages, the Australian language context is different and not generally represented in training corpa, nor should they be. However fine tuning, RAGS or prompts may help these models understand our languages.

Hence the research aims are to:

1. Evaluate perceived value of AI tools amongst community language learners and teachers and improve this through workshoping suitable models.

2. Estimate the value of LLM models fine tuned or supported with a RAG for these languages and consider the amount and type of local material needed to provide useful models

3. Evaluate the optimum pipeline for various translations formats between languages and media (eg text and speech),

4. Provide community evaluation criteria for such models that can be used to support future improvement of models

1.4 Projects

In all the work described here the data is not available for further testing due to cultural protocols. This work is done locally with each language to provide community-governed infrastructure and tools. Hence using cloud models works against that purpose, even though such hosted inference models were useful during development.

This paper covers two types of projects. One area is to translate from Text-to-Speech (TTS) and Speech-to-Text (STT). The other area is translation models provided as an API tutor based on the written material, supported by the dictionary and grammatical analysis done on these languages.

When developed together on one language these could form a pipeline for speech-to-speech (STS) translations. However, this work will require large amounts of interpretation between cultures and the complex details of grammatical structures in languages which poses technical challenges for Natural Language Processing (NLP) models, which are optimised for more widely spoken languages like English. Overcoming these barriers requires a combination of technical innovation, community co-design, and respect for linguistic diversity.

The audio projects were by university staff working with local community languages and running workshops with the speakers; and the text projects by a NSW language centre and a community language speaker who has fine-tuned his own models. The audio pipeline is also being replicated with the NSW language centre.

The first project provided a TTS system in one strong language and STT in another one from the northern Northern Territory as these were the models requested by the speakers after the first workshops demonstrating what AI models can do. This was run by a university team led by a First Nations academic working off country.

The on-country work was run by an academic working with the Gumbaynggirr speakers in northern NSW and a second project by a Yamaji man from the Gascoyne region of Western Australia through his mother and Māori (Ngāpuhi) through his father. This community author is an independent developer building language tools with personal resources. The project was based on Wajarri and provides a Wajarri Wangga Android dictionary and learning application, with several model-adaptation and corpus-building attempts.

The goal of these projects are systems that can use the available language material to produce useful oral and text translations, explanations and creative learning material, rather than simply return dictionary words. .

2 Related Work

2.1 Bilingual schooling

Much of the language reclamation work is done in language centres at universities or within government funded community language centres. This provides a way of collecting the material that has been archived and then sharing back to community to start the speaking or encouraging those outside the community to learn the culture and language of those they now live with. The majority of language learning in Australia is now carried out in schools, so while many are not comfortable with language work being located in schooling, it does provide an avenue for learning. The bilingual programs, while mostly dismantled in 2008 in the NT, are again becoming a significant focus of community and language workers.

The text projects are based on the archival text material, collected in dictionaries in the first instance to provide a standardised spelling and translation. However this is merely one version of the large knowledge system. Audio material is collected from the readings of old books or new material related to health and wellbeing support in language.

For STT there is the opportunity to support learners to give examples of what their spoken expressions sound like to a listener, but also we have First Nations working as interpreters, such as in courts and hospitals, who are also asked to provide translations for government and company messages. These are requested as text, however most are not strong in writing in their language. Furthermore, many of these speakers could work in schools, however again written skills are required. If we can provide rough translations for them to then improve, this has been seen as a useful approach in community workshops.

Finally the meaning of words, how to construct into sentence remains a crucial part of learning, or for our youth relearning our languages. The speech experiments and the Wajarri developer case address complementary stages of language technology. The former examine learning from recordings and transcripts; the latter show the work required to assemble textual evidence, learner-facing applications, evaluation and ongoing operation. Across both, a working demonstration is a starting point for deciding usefulness, and providing permissions and the new resources needed to continue to improve the models.

For oral projects we have a language agnostic mobile app Dalang-Gathering (Charles Darwin University, n.d.) that allows speakers to upload the words and sentences they want to record, and then they can individually record these, linked to their written form. These are stored on a secure cloud database to be downloaded and used for future model training.

2.2 Augmented intelligence (AI)

For TTS, models now use deep neural networks to extract and reuse the complex mapping between text and speech. This was previously done by handcrafted rules and features. State-of-the-art systems typically employ a multi-stage architecture with various parameters to adjust learning and performance. The system used consists of a text analysis module that codes the text for an acoustic model which predicts the properties of the speech.

A vocoder model generates the waveform output through predicting the frequency content. Neural TTS models, such as Speech2, WaveNet, Tacotron, and FastSpeech2, have significantly improved the quality of synthesised speech by directly modelling raw waveforms or generating them from acoustic analysis like mel-spectrograms. This allows for more natural-sounding and expressive speech than traditional methods. Work on Canadian languages by Pine et al. (2025) provided a two-stage system using a feature prediction network based on FastSpeech2 and and a vocoder based on iSTFTNet, see `https://docs.everyvoice.ca/stable/`. Like this system we hope to generalise between languages by providing a pipeline to tailor the underlying model to local data.

The adoption of STT technology for languages with limited digital resources can provide digital inclusion and linguistic fairness to these languages by reducing the digital gap, providing a means for the digital preservation and respect for First Nations languages. There have been efforts to support Low Resourced Languages with STT models, such as Extremely Large Speech Recognition (XLSR) models (Conneau et al., 2020) with application across 53 languages in various linguistic environments. Further exploring the XLSR models Baevski et al. (2020) described the Wav2Vec model used in this paper for audio work, demonstrating that by using a pre-training phase of self-supervised learning the required training data can be greatly reduced. Research such as that by Khurana et al. (2024) have applied cross lingual transfer from high to low-

resource language in XLSR and suggest that the model can be effective for transcribing languages which have limited digital resources.

The tools we develop are intended as augmented support for learners and teachers as they are prone to error when built on minimal data so are not highly reliable. This is particularly the case in text translation. The major challenge is this scarcity of semantically annotated data, which is essential for training models to grasp the specific nuances of the language (et. al, 2026).

We are working with audio from the stronger spoken languages of the Northern Territory in Australia and consider then the application of these models to other languages being reclaimed. The written work are based on separate community projects using their dictionaries as accepted by community.

2.3 Text-to-speech

For text-to-speech in low resource languages multilingual models have been proposed. These are models trained on data from multiple languages, aiming to leverage shared linguistic features to improve performance on a low-resource target language. However, the complexity of multi-lingual training and the diversity of linguistic features can pose challenges (Byambadorj et al. 2021).

Instead we considered transfer learning. This approach involves using models pre-trained on large, high-resource language datasets and then fine-tuning them on smaller, target low-resource language datasets (Azizah et al., 2020). This considered the new phonemes to be generated for the languages to be accurately spoken by computer. A TTS method which has provided greater opportunity is using models pre-trained on large, high-resource language datasets and then fine-tuning them on smaller, target low-resource language datasets. This study utilises this method by employing a Tacotron 2 model initially trained on the LJ Speech dataset for English, which we fine-tuned using 15 and 19 minutes of Yolngu language data of 2 different speakers. This strategy effectively addresses the challenge of data scarcity common to low-resource languages (Azizah et al., 2020).

Tacotron 2 utilises an advanced encoder-decoder architecture, with attention mechanisms, that has proven effective in generating high-fidelity, natural sounding speech. Hence it was chosen for its superior ability to capture the natural intonation and nuanced linguistic elements needed for First Nations dialects

To evaluate if a computerised voice that our models generate would be accepted, and also to remove the concern of using known voices of deceased people we used model agnostic meta-learning. This adapts TTS models to new speaker characteristics. By requiring fewer steps for adaptation, it enables efficient voice cloning and personalisation but high computational resources (Huang et al., 2022). We used XTTS-v2 for its ability to model new voices using only a 6-second audio reference. It achieves this zero-shot voice cloning with a GPT-style autoregressive model combined with a vector quantized variational autoencoder. By requiring fewer steps for adaptation, it enables efficient voice cloning and personalisation. While it maintains high-quality voice outputs, the method may demand substantial computational resources and specialised expertise for effective implementation (Huang et al., 2022).

2.4 Speech-to-text

The STT models traditionally considered are Deep Bidirectional LSTM, which have shown noteworthy performance in multilingual contexts (Graves et al., 2013), and Transformer models, which have been applied to tasks involving sequence-to-sequence translations, and now used frequently in machine translation (Sun, 2024).

When it comes to low-resource languages there are more challenges, as they require models that are both robust and adaptable. This is where the XLSR-Wav2Vec2 model provides support. Developed as a large, pre-trained language model, XLSR-Wav2Vec2 has a multilingual foundation, trained on vast amounts of data spanning multiple languages (Baevski et al., 2020). Its adaptability, especially its capacity to be finetuned for specific linguistic nuances, makes it an ideal candidate for low-resource languages (Artetxe and Schwenk, 2019).

Again the chosen approach was to use fine-tuning as a crucial step when adapting pre-trained models like XLSR-Wav2Vec2 to specific tasks or datasets. It ensures that the model can effectively handle the nuances of the new language (Howard and Ruder, 2018).

2.5 Dictionary translations

The Gumbaynggirr system uses the llama inference engine to run a Qwen3.6-35B model (see `https://arpanlabs.com/gumbbot/networking.png`).

The dictionary data was extracted using Qwen and has since been updated so the data will be regenerated using the pipeline developed. Qwen (Tongyi Qianwen) are open-weight large language and multimodal models developed by Alibaba Cloud. They are commonly used for natural language processing, coding, reasoning, and visual understanding

The Wajarri project used a similar approach to extracting the dictionary data and developed an inference model from a larger cloud Qwen-122B system but now uses a local Qwen-27B configuration fine-tuned with LoRA.

For the Wajarri project 2,620 supervised training rows and 288 validation rows were recorded. These counts establish prepared material but did not provide a fluent trained model. The current public text generator was developed from earlier adaptation experiments plus a deployed fine-tuned retrieval component.

Early work has included synthetic training data, morphology analysis, model adaptation and retrieval. After some runs the community author found that fabricated forms, repetitive constructions and uncertain language quality repeatedly forced revision. In each case the pipeline had to be regularly rebuilt around the source dictionary, with provenance and admission checks.

As it was unclear whether the training data was usable, it was used to translate the Bible, Quran and other documents as a source of connected language. The Genesis data contains 1,533 verses using token-attestation and model-fidelity checks. Linguistic feedback nevertheless identified grammatical and contextual errors.

The public tutor uses dictionary, example and grammar tools to support generated responses. The checked dictionary edition contains 2,292 entry records and 1,575 example sentences. The archive's 105 non-empty pseudonymous visitor identifiers were collected on usage. Small composition replays showed some improvement, while longer requests still failed or exhausted their budgets on the model.

3 Methods

In this section we describe the datasets and methods we used. These are not comprehensive, merely an analysis of the approach we took and why, and the success of these models to date.

3.1 Datasets

We are using a dataset consisting of samples in various Australian Aboriginal languages, each one forming the corpus of a separate project. Each text sample represents a transcription of the spoken audio recording. The TTS and STT were carried out on different languages for different needs. For text based translations we use the dictionary accepted as authoritive by the community. We describe here the projects completed so far.

3.2 Text-to-speech

We adapted transfer using a dataset consisting of 488 audio samples of one language. These audio samples were resampled at 22,050Hz down from 44,100 Hz. Each audio clip was limited to a maximum length of 10 seconds, similar to the LJ Speech public dataset in English. We used a Tacotron 2 model pre-trained on the LJ Speech dataset, a 24-hour corpus of English-language data. We then fined tuned the model using 15 minutes of language data and further followed it with an additional 19 minutes of audio.

To further improve the model's performance, we extended the dataset to 1588 audio samples totaling 98 minutes, combining recordings from three speakers across various dialects of the language. To account for specific sounds of the language, we extended the English character set supported by the model by adding symbols not used in English.

We applied early stopping based on the validation set and trained the model for 54 epochs, at which point validation performance stabilised. Training progress was tracked using training vs validation loss graphs for different dataset configurations Audio quality is measured using encoder-decoder time step values, which range from 0.0 to 1.0. These values help us see how well the input text aligns with the generated speech. When these values form a diagonal line (see Figure 1), it indicates that the model is capturing the timing and structure of the text, resulting in clear and high-quality sound with minimal distortion. Most distortions found in early work were removed using the full 98 mins of mixed-speaker audio. When the models have been fine-tuned, we tested then with human speakers who provided the final evaluation of the TTS models. Initially with interpreters of a strong language of the NT, then with Gumbayngirr in their language centre.

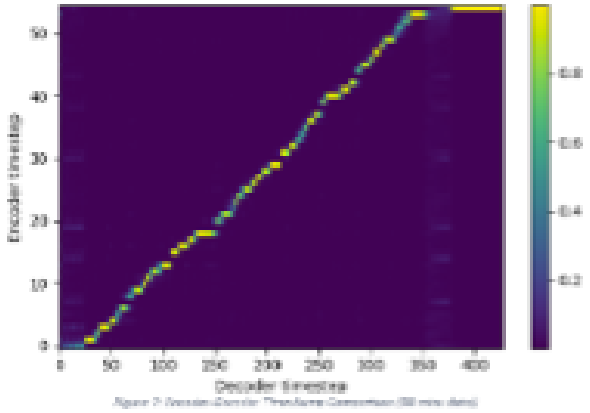

Figure 1: Audio Quality for TTS output

3.3 Speech-to-text

For audio data pre-processing in STT we used a different data set of 1,758 mp3 audio files converted into 16kHz sampling rate, provided by the language speakers. We used a 16 kHz sampling rate to maintain a consistent data structure and ensure compatibility with the speech-to-text model.

For preprocessing the transcription data contained unwanted spaces, numbers and character errors which were cleaned. The training set contains 1,582 pairs (90% of the dataset) of audio and text, and the validation and test set each contained 88 audio-text pairs (5% of the dataset each). To improve computational efficiency and ensure similar length sequences are processed together, data was grouped by length (Howard and Ruder, 2018). An essential parameter, the per-device training batch size, was defined to regulate the number of training samples processed simultaneously, balancing between computational constraints and training stability. The evaluation strategy was set to steps, promoting regular assessment throughout the training process as suggested in image processing (Wang and Perez, 2017).

With the training epochs set to 80, the model underwent iterative learning cycles, refining its weights and biases. To further optimise training, 16-bit training was enabled, along with gradient check-pointing, streamlining the computation process. Other hyperparameters, such as learning rate, weight decay, and warmup steps, were calibrated to ensure a steady and adaptive learning pace. Given the extensive training involved, checkpoints, evaluations, and logs were set at regular intervals of 500 steps.

This system was tested with language tutors at the university for the sstong NT language and then re-used the process for Gumbaynggirr tutors at their language centre.

3.4 Translation

In translation the work is to extract as much of the language information from the dictionary, which are often organised under different metadata, depending on the linguist. The dictionary is extracted as a json file and other resources added to form the RAG with a tailored retrieval and answer prompts to restrict hallucinations. The Gumbaynggirr model is set up to provide answers to translations tasks in either direction, but is forced to be strongly tied to existing example data from the dictionary and requires more linguistic patterns to extend to new sentences. It is now being tested by speakers as to the value of the system.

In Wajarri project the computer capacity (a home laptop) has limited the work to a small model, however the size of the model was not the sole explanation for inaccuracies in translation. This was tested by the community member working on this project and the linguist working with their language. An early comparison between local answers and archived cloud answers after internet failure made cloud not an option. For eleven comparable answers, judges of the model preferred cloud grounding in seven cases and local grounding in four; preferences for teaching data preferred the local model in seven, cloud in three, with one tied.

Adjustment of the model using grammatical features and specific language knowledge was used to correct the model however these manual efforts merely led to errors in other examples. The issue remains that the resources are not sufficient for reliable translation. Hence we look to assess if it still can be used as guidance for language teachers and learners.

4 Results

This work was done with different languages in separate projects. The significant factors came from the outcomes of the second stage where we demonstrated the models in speakers' workshops or presentations. In each case the assessments were made by those who requested the models for their specific use. For TTS two tutors teaching the language evaluated the model for 15 minutes. For STT we had four interpreters evaluating the model together in the strong language, and two teachers evaluating separately in a reclaimed language. These have

shown that a small number of text-speech pairs can generate a system that incorporates some of the language characteristics such as new phonemes and language prosody. The issues or outcomes that we found so far are summarised here to encourage others to venture into sharing with different community members the opportunities in this field.

We are gaining a greater understanding of how these Transformer AI models work. The STT models lose their knowledge of English when fine-tuned (chronic forgetting). When provided with new sounds but with limited data, the speech can become stuttered. As found by others experimenting in this area (Pine et al., 2025), encoding spaces or removing them can reduce this error. As there are different ways to introduce new data to a model, through fine-tuning, RAGS or prompts, we are now experimenting to find the optimum pipeline to adapt models to new low resourced languages.

For the audio projects we used data from two different strongly spoken languages which are from the Northern Territory, Australia. The data comprises single written words and sentences with their audio. This is not sufficient data or an accurate model available for sharing but was sufficient to verify to community that such a model can be achieved with the data type used. The two models have been developed separately and the results in this paper are mostly from their first tests.

At present the data cannot be shared due to cultural protocols, however the results allow us to return to the community to consider the value of these models and if they warrant further recordings. We provide here some of the feedback from these sessions.

4.1 Text-to-speech

The results of the system are shown in Table 1. This downward trend in MSE during the training phase indicates that the model's predictions are becoming increasingly accurate over time. Close monitoring of the encoder and decoder activities within the Tacotron 2 model was performed during the training phase. This provided critical insights into the model's performance, allowing for fine-tuned adjustments.

To further analyse the encoder-decoder alignment and performance, time-stamp comparisons were performed for different datasets. These comparisons provide insight into how well the model captures the timing and structure of the input text and audio, as shown in Figure 1using the full 98 minutes of data from different speakers. The encoder translates the input text into a condensed numerical representation that the decoder can interpret. The decoder then utilises this representation to generate a spectrogram, which is converted into audible speech.

Table 1: Mean Squared Error Over Validation on Training

Iterations	MSE Value
24	0.231358
264	0.163634
504	0.165322
744	0.167714
984	0.180726
1224	0.184559

Testing: Human testing revealed that, while the initial results were promising, certain sounds, particularly 'ng', were distorted. Further analysis showed that these distortions were linked to pauses in the original recordings. Data cleaning by removing unnecessary pauses, enables the model to produce clearer and more accurate speech outputs. This process highlighted the critical role of high-quality, clean data in training effective TTS systems.

Community Response The TTS model was demonstrated to two of the tutors in this language at the university. They are teaching as well as doing their research. They felt the output of the speech generation was generally accurate, however some of the 'new' sounds (not in the English base language) were not produced well. However the "sound" of the words in terms of prosody and accent was much more acceptable than older systems which sound computerised. The resemblance to a human voice is an important factor for community acceptance.

However an issue arose around protocols over the use of someone's voice. The speaker in the recordings was related to the tutors and as language is the expression of knowledge, that knowledge is aligned to the speaker. They found hearing the voice speaking words that were not actually spoken by that person was 'spooky'. There is further protocols to be considered for voice material of a deceased person as our teachers pass on. The voice and images of the deceased are upsetting for Aboriginal people as they call them back to this

world.

When using different voices in the strong NT language model, the speakers were talking about very different domains, hence the spoken form matched the voice of the person who recorded most in that domain. There are however technical solutions available as used in our subsequent work with Gumbaynggirr.

Since this work was done, we have had further requests from community-based school teachers in a related dialect. While teachers are present in high numbers in these remote schools, it is still seen of advantage to provide on-the-spot feedback on students' writing with an audio version, since these are audio languages. However we have been told we cannot use the resources between languages, not even between dialects, and that the new dialect will need to re-record their own material.

4.2 Speech-to-text

There is a decrease in Training Loss until the 2000th step and stabilised afterwards as shown in Figure 2, which hinted that the model was approaching towards its performance ceiling with our dataset.

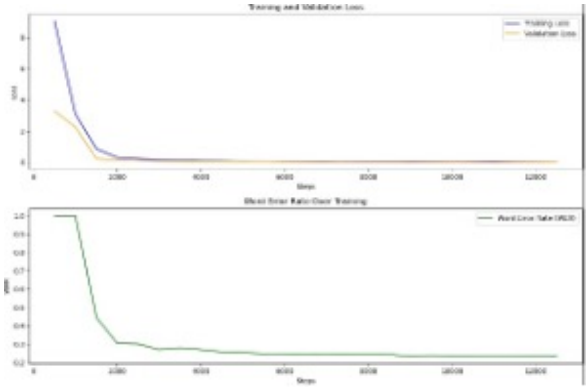

Figure 2: Audio Quality for STT output

Word Error Rate (WER) metrics used in speech recognition show the model's refinement over the training steps. Beginning at a value of 1.0000, its consistent decline culminating in a WER of 0.24 between the 8500th and 9000th steps indicated the model's sharpened accuracy, resonating with findings in contemporary ASR research (Baevski et al., 2020) . With reductions of 99.79% and 99.41% in Training and Validation Loss respectively, and a 76.68% enhancement in WER, the model's efficacy in transcribing was evident (See Figure 2). This system is now configured for user testing with our speakers rather than machine testing as we are evaluating the pipeline rather than one data set.

Community Response The STT results were not as successful as the TTS in that we used a purely Wav2Vec model with no post-processing. Hence the spaces used in spoken wording did not necessarily match the text word segmentation and the difference in the voice of the speaker to that used in training the model has significant effect on the human evaluated WER.

While workshop participants found the output amusing, and thus appreciated the limits of these machine models, the approach can provide some of the assistance that was originally requested for transcribing. We are seeking more data now to improve the model outputs such as setting up a grammar dictionary for a RAG

We are dealing with languages that were never written until English people came to Australia and yet insist the written form be used for communication. Hence, we aim to support local speakers to fit within that culture to survive but also decreasing the reliance on text in language learning.

We have seen that the form of the language is important. As well as old forms that people want to retain, we have different spelling created by different linguists working with languages and these can create conflicts that slow the sharing of resources. In future by developing STS translations, we can avoid the disagreements over the intermediary text stage and have language to english developed audio to audio

4.3 Translation

These system are the most amenable to local community development as we have the dictionaries but often not much audio recorded and transcribed. With the support in AI for programming, we are finding community teachers are setting up their own models and running dictionary analysis to produce online translations tools. However the variation in dictionary formats and individual entries makes it hard to collect the information in a format to be used by the AI models, so we use AI to clean the data.

The Gumbaynggirr model is still just a basic tool that stays with the existing sentence examples from the dictionary and avoids hallucinations. The next approach will be to consider the grammatical knowledge that can be incorporated into the translator. Similarly for Wajarri Wangga the model provides a complementary dictionary and learning interface, including flashcards and pronunciation

material with a history and provenance of the language examples, but its ability to deal with new sentences is limited.

Community Response Feedback from Wajarri linguists identified grammatical and contextual errors, including verbal morphology and selection of English senses. Plainer English source phrasing was also suggested. This showed that dictionary checks alone could not establish sentence-level correctness, and that inconsistent terminology was only one of the outstanding problems.

Added to this was the problem of timeouts during translations as the model was running on a home laptop. The computing resources available for this work is also limited.

5 Discussion and Conclusion

Text-to-Speech and Speech-to-Text systems may not be perfect but they provide some support in conversion between text and speech and are clearly possible with the new transformer systems and offer opportunities to support teachers and learners of low resources languages and interpreters. To create a full translation system will require more accurate models and these project continue to work with communities for approval to collect more data.

The described architecture and its associated capabilities are particularly advantageous when adapting TTS technology for low resourced languages. The model's proficiency in capturing the natural intonation and nuanced linguistic elements of speech aligns well with the project's dual objectives of utility and cultural preservation. While the models show promising learning capabilities, the later stages of the training indicate challenges related to overfitting.

The significant improvements in Training Loss, Validation Loss, and Word Error Rate (WER) in STT confirmed the adaptability and efficacy of the XLSR-Wav2Vec2 model. Furthermore, the observed model converged around 5000 steps with low loss figures.

For the translation system the errors are clearly visible to anyone skilled in the language, hence the work can only be a guide, not an authority. However, communities recognised their languages were being valued by incorporating them respectfully into modern technology.

Future needs are sustained developer time on all projects, reliable computer capacity, paid speaker and linguistic participation, and ongoing workshops with communities to decide what language can be taught or admitted to training. Exhaustive review is itself a community workload. The governance proposals and organisational discussions over the data continue and annotated recordings, transcripts and permission for the intended use would be needed for any speech project to continue that work.

These machine learning models are designed to support the revitalisation of these languages: a text-to-speech model based on Tacotron 2 and a speech-to-text model leveraging Wav2Vec 2.0. The TTS model is adapted to the language vocabulary and generates spoken language from text. The STT model enables the transcription of spoken language word by word. Our experiments show that these systems perform relatively well despite the limited annotated data and the unique phonetic and structural characteristics of the language.

The final evaluation of the TTS model developed in this paper was to use it to provide oral examples of an unrecorded language from the east coast of Australia. Existing phrases in the language are written only, and the output was similar to what is now the language taught but provided an authentic accent for learners to hear. Hence in future the TTS model may be considered for application across languages with similar phoneme sets. As there are still many sounds in the phonology of Aboriginal languages which are not in the other commonly spoken languages, we will require more data to enable the models to understand these and provide correct output in both directions between speech and text media. We would like to use the data from the strong languages to support languages which were never recorded before their use was mostly stopped. However, such a proposal must be approached with clear outlines of what can be done and who has access, as it involves sharing someone's voice data.

At present the models are trained separately on each language to retain sovereignty of the speakers. While the initial results demonstrate the potential, further work is required to refine the models and adapt them to a broader range of languages and dialects.

Future directions include integrating larger and more diverse datasets provided by community who see value in the models developed so far, and continuing our collaborations to ensure cultural appropriateness and usability. The work to collect audio-text pairs or transcribing recordings requires

dedicated language experts. We hope the experiences will encourage more to be involved when they see the output in the improvements to their models. Finally, we hope by gaining approval to share these models as examples, we will encourage more languages to consider such tools for their use.

Improving text translation models will include using intermediary languages like Kriol for the NT which provides the interpretative language from English to First Nations language structure and context. In Wajarri the work on text needs much more computing resources and advice on optimal pipelines, but shows what can be done by community based First Nations who are keen on investigating the use of technology to benefit their language reclamation.

Acknowledgments

The audio work was under CDU HREC approval id H24072. The dictionary projects are run by language centres and speakers with their community approval and the process is reported by the community themselves.

References

Mikel Artetxe and Holger Schwenk. 2019. Massively multilingual sentence embeddings for zero-shot cross-lingual transfer and beyond. *Transactions of the Association for Computational Linguistics*, 7:597–610.

Australian Research Data Commons (ARDC). 2024. Care principles. https://ardc.edu.au.

Kurniawati Azizah, Mirna Adriani, and Wisnu Jatmiko. 2020. Hierarchical transfer learning for multilingual, multi-speaker, and style transfer dnn-based tts on low-resource languages. *IEEE Access*, 8:179798–179812.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Preprint*, arXiv:2006.11477.

Charles Darwin University. n.d. Dalang language gathering app. https://github.com/catkutay/Dalang-Gathering-App. Private access with permission, needed to allocate database.

Commonwealth of Australia. 2020. *National Indigenous Languages Report*. Australian Institute of Aboriginal and Torres Strait Islander Studies (AIATSIS), Canberra.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. *Preprint*, arXiv:1911.02116.

Seb Dianati and Reuben Bolt. 2025. Embedding indigenous knowledge in the curriculum: Principles and strategies across the australian higher education sector. *The Australian Journal of Indigenous Education*, 54(1).

Tianyang Zhong et. al. 2026. Opportunities and challenges of large language models for low-resource languages in humanities research. *Preprint*, arXiv:2412.04497.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. 2013. Speech recognition with deep recurrent neural networks. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6645–6649.

Camiel Hamans, Hein van der Voort, Miguel Oliveira, Jr, and Katia Valdenilson. 2024. Economic value of endangered languages. *Cadernos de Linguística*, 5(e723).

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. *Preprint*, arXiv:1801.06146.

Sung-Feng Huang, Chyi-Jiunn Lin, Da-Rong Liu, Yi-Chen Chen, and Hung-yi Lee. 2022. Meta-tts: Meta-learning for few-shot speaker adaptive text-to-speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:1558–1571.

Sameer Khurana, Nauman Dawalatabad, Antoine Laurent, Luis Vicente, Pablo Gimeno, Victoria Mingote, Jonathan Le Roux, James Glass, Mitsubishi Electric, and Research Lab. 2024. Cross-lingual transfer learning for low-resource speech translation. *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, pages 670–674.

Aidan Pine, Erica Cooper, David Guzmán, Eric Joanis, Anna Kazantseva, Ross Krekoski, Roland Kuhn, Samuel Larkin, Patrick Littell, Delaney Lothian, Akwiratékha' Martin, Korin Richmond, Marc Tessier, Cassia Valentini-Botinhao, Dan Wells, and Junichi Yamagishi. 2025. Speech generation for indigenous language education. *Computer Speech and Language*, 90:101723.

S Romaine and D. Nettle. 2000. *Vanishing Voices: The Extinction of the World's Languages*. Oxford University Press, New York.

Y Sun. 2024. The evolution of transformer models from unidirectional to bidirectional in natural language processing. *Applied and Computational Engineering*, pages 281–289.

X Wang and L. Perez. 2017. The effectiveness of data augmentation in image classification using deep learning. *Preprint*, arXiv:1712.04621.