Proceedings of the LSA

Speech-based generative AI for low resource languages: Promises and pitfalls

Claire Bowern & Alessio Tosolini*

Abstract. Existing collections for spoken languages have numerous barriers with which speech technology can assist. In sharing forums, Indigenous Australian community members often raise orthography as a barrier in language work. For some Native American communities, text to speech systems (TTS) have proved useful. We compared TTS model outputs for one Australian language, and in doing so, ended up creating a deepfake of an elder. Existing guidelines do not adequately cover the possibilities and concerns raised by this area of generative AI and protocols for the use of archival materials for this type of speech technology are urgently needed. We suggest some preliminary guidelines.

Keywords. Speech technology; text-to-speech; ethics; endangered languages; phonetics; voice synthesis; Indigenous AI

1. Introduction. Indigenous language activists repeatedly express frustration at the barriers and the amount of ancillary information they need to acquire, in order to do language work and connect with language materials. A language activist from northwest Australia expressed it this way:[1]

> All I wanted to do was learn my language, but to do that I had to get a degree in linguistics and learn a whole bunch of other stuff that wasn't relevant, it became my entire life.

That is, the barrier to language learning for many Indigenous languages is much higher than it is for better studied and resourced languages: in terms of access to low-stakes activities such as Duolingo, accessibility and availability of information about the language and in the language, and ways to help emerging language users continue learning on their own.

Text and audio for spoken languages is one such nexus point. For example, learner's guides often provide a pronunciation guide that is very abstract (or relies on knowledge of other languages, e.g. "trill the *r* like in Spanish"; Bowern et al. 2007). Providing audio information directly in the dictionary would be helpful, just as videos of signs are more useful than static pictures for signed language dictionaries. However, making such materials is a high bar, especially for languages without extensive existing resources or where learners are not able to get much exposure to the language directly—that is, in the same contexts where language activists identify problems and frustrations in the first place.[2] This is therefore one area where audio technology may advance language access goals.

* Acknowledgements. We acknowledge Yan-nhangu community members: yapa-miṯṯji ga yolŋu Nhaŋu-mirr. CB acknowledges NSF grant BCS-1423711 and ELDP FTG-0011 (2004–2007); opinions expressed here do not reflect those of the NSF. Authors: Claire Bowern, Yale University (claire.bowern@yale.edu) & Alessio Tosolini, McGill University (alessio.tosolini@mail.mcgill.ca).

[1] The quote is a paraphrase from a longer discussion at a language activism meeting in 2021.

[2] Even if audio recordings exist for particular words, they might not be usable for dictionaries. For example, in extracting words from field recordings for an electronic version of the Bardi dictionary (Aklif 1999), we found that many words had not been recorded in isolation or in citation form, making the word difficult to isolate. Other items had background noise or multiple speakers talking simultaneously.

Other ways where computational tools would aid in language reclamation include facilitating access to the large amount of material that exists only as written text. Synthetic audio for written materials does not provide the same connections that hearing the voices of the original storytellers does, but it can assist in understanding. For example, in 2008 the first author read Bardi stories aloud and burned CDs for Bardi community members in order to better share narratives from the 1930s that existed only in a bespoke orthography and difficult-to-read handwriting.

Another use case is self study and feedback. New speakers often want to practice and gain confidence in the language. Having access to audio recordings that new and emerging speakers can model can be helpful. Again, however, creating such resources manually is extremely time-consuming. Recorded speech might allow numerous replays, well beyond the repetition tolerance for a speaker.

One way to address this bottleneck is to use speech synthesis. It's now possible to create naturalistic synthetic voices with only a few hours of data. This makes them tempting for language reclamation and support projects. However, using such methods can raise numerous ethical concerns, crossing over from synthetic voice to deepfake. In this paper, we describe our experiences in creating text-to-speech (TTS) technology for an Indigenous Australian language of Northern Australia, and how a project that was set up well within regular parameters very quickly ended up in problematic ethical waters.[3]

Our goal in this paper is therefore to describe what we did, what we see the ethical issues as being, and suggest ways to avoid this happening. This paper is about what led to that work and how not to repeat our mistakes. It is urgent and important to address these problems because of how easy it is to implement the technology, and because of the potential for harm.

In the discussion following the presentation of this work at the LSA's 2026 Annual Meeting, it became clear that community norms around these issues differ substantially. We emphasize that these issues particularly arise in our case because of two points. Firstly, we are working with archival materials where the original contributors are all deceased, and therefore are not able to give consent or discuss permissions. Secondly, our use case involves generalizing across languages, disrupting connections between languages and identifiable individuals. These points might not arise in other contexts, and indeed, similar types of synthetic voice creation are increasingly common in language learning contexts. This paper does not claim that that language activists should not create these models and use them to advance their own language goals. Rather, we highlight the ethical concerns we have encountered for our case-study of speech-based generative AI for Indigenous Australian languages.

2. Background. As is well known, written representations of spoken minority and Indigenous languages have numerous barriers to use. Materials may be written in unfamiliar orthographies (or orthographies with unfamiliar conventions) which may be challenging for new learners, who have to learn conventions in the abstract before applying them. Literacy skills from one language are transferable to another with practice, but individuals learning languages outside of a regular classroom often lack access to that training. If learning primarily from source materials not created for language learning (such as reference grammars or field notes), learners may be presented with multiple orthographies or conventions.[4] Such materials are not made to give learners feed-

[3] We have not released any of the synthetic voices we made and will not do so.

[4] We know of no estimates for the number of languages that have genuine pedagogical materials; `glottolog.org`, for example, tracks types of materials aimed at linguists but not materials aimed at communities. Anecdotally, how-

back on pronunciation. Even in contexts where pronunciation norms vary widely, having access to the auditory targets represented by orthographic conventions is helpful.

2.1. USING TEXT TO SPEECH. Text to Speech technology (TTS) can address some of these issues. By creating audio representations of written materials, learners and other language users can get more practice with, and more exposure to, the phonology of the language. These audio representations not only provide auditory feedback to help with decoding and practice, but also provide a way to get a better understanding of how the orthography works. On the materials creation side, using TTS makes for rapid creation of speech-based materials without placing demands on speakers' time. This is important where it's not feasible to make large-scale spoken word materials due to time, cost, or availability of speakers. It also allows for the creation of audio for items that were originally recorded in print but not on audio, with the added benefit of being adaptable across languages with similar inventories.

Additionally, training models on synthetic speech data has also been shown to improve accuracy in certain low-resource automatic speech recognition settings for Hungarian (Mengke et al. 2026) and Seneca (Thai et al. 2019). Many under-resourced languages have an abundance of unlabeled or unannotated data that is largely unusable for model training due to the lack of corresponding audio, transcriptions, or glosses. By automatically generating audio files for preexisting unlabeled textual data, TTS can augment the amount of training data available for speech technology that trains on pairs of text and audio files, such as automatic speech recognition. Although the quality of generated audio is generally lower than that of recorded speech, these audio data sources may be especially helpful in expanding the distribution of words that other acoustic models see, broadening the scope of applications of such models.

There are increasing efforts to create speech synthesis technology for low-resource languages, with a surge in technology for North American Indigenous languages in particular. Intelligible non-neural models have been developed for Navajo (Whitman et al. 1997) and Plains Cree (Harrigan et al. 2019). More recently, teams developing TTS models for Border Lakes Ojibwe (Hammerly et al. 2023) and Mundari (Gumma et al. 2024) have found success using non-autoregressive models such as VITS (Kim et al. 2021) trained using the Coqui AI framework (Gölge & Team 2025). Another successful instance of non-autoregressive models used for TTS development include a series of models based on FastSpeech2 (Ren et al. 2022) developed for Kanyenkha, SENĆOŦEN, and nłhiyawłwin (Pine et al. 2025), with an open-source library being developed and at the pre-beta state of development as of June 2025. Finally, a multilingual and multispeaker TTS system using the non-autogressive MatchaTTS (Mehta et al. 2024) architecture has been used for Ojibwe, Mikmaq, and Maliseet (Wang et al. 2025).

2.2. HOW TTS WORKS. Modern speech synthesis generally works by combining two component: an *acoustic model* which converts input text to a series of acoustic features, and a *vocoder* which takes the acoustic features and converts them to waveforms. There is great variety in the architectures of acoustic models, such as whether acoustic models are autoregressive, meaning generation of a given audio frame is conditioned on the previously generated audio frames, or non-autoregressive, meaning that the entire output is generated in parallel. Additionally, models may include attention mechanisms that are data-hungry, requiring large amounts of training data to yield good results (Pine et al. 2022). Autoregressive acoustic models, such as `Wavenet`

ever, the number of learners of Indigenous languages in the US and Australia who first encounter their languages through materials made for purposes other than pedagogical ones is substantial.

(Oord et al. 2016), `Tacotron` (Wang et al. 2017), and `Tacotron 2` (Shen et al. 2018) will not be discussed extensively in this paper since they tend to require larger amounts of data to successfully generate audio and thus are not suitable in the low-resource settings we are interested in (Pine et al. 2022). In this paper, we therefore opt to use MatchaTTS (Mehta et al. 2024) since it's non-autoregressive and has been used successfully for other low-resource languages (Wang et al. 2025). We compare this to a fully synthetic non-neural TTS app, eSpeak-NG.

In addition to the locally run models referenced here, there are numerous commercial products which require users to upload data, which may be used to further train the commercial models. There are also products, such as Pincel's avatar creation, which aim to create animated avatars of single individuals. We only consider options that release no data to third parties. Such apps also have chatbot components, which are not part of this test at all (that is, at no stage were we interested in creating a chatbot using the language).

2.3. THE CURRENT PROJECT. The current project began with the first author's work with the Kullilli language reclamation project Kullilli Ngulkana.[5] Materials for this language project include audio recordings from several Kullilli speakers but the vast majority of sentence data are printed in McDonald & Wurm (1979) and Holmer (1988) and were probably never recorded. The dictionary we compiled has roughly 800 headwords but fewer than half of them can be found in the audio sources. Community members wanted to be able to hear the language and agreed with preliminary investigation of whether computational approaches would likely be feasible.

A TTS model created for Kullilli may well be usable by other Australian communities too. 80% of Australian languages have similar phonological inventories and stress placement (Round 2023) implying that resources created for one language might be transferable to another. We decided to use Yan-nhangu (Yolngu, Pama-Nyungan; Baymarrwaŋa et al. 2005; Bowern 2023) speech data to make an exploratory and prototype TTS system.[6] We know the circumstances of recording as they were made by the first author. More importantly, the Yan-nhangu elders who made the recordings had given permission for secondary uses which included projects beyond Yan-nhangu (for which see Section 3.2 below). There are other syustems that use a single model for Australian Indigenous languages. WebMAUS forced alignment, for example, has an "Australian" option, a general model that is aimed at Australian languages in general.[7]

3. Methods and Data.

3.1. DATA DETAILS. For training the TTS models, we used field recordings made by the first author with Yan-nhangu speakers. These materials have over 1,000 translated sentences and are relatively clear recordings, made in quiet indoor and outdoor locations with relatively little background noise.

The Yan-nhangu language team made recordings over the period 2004–2007. This fieldwork was funded by the ELDP and was initiated by a community member with the aims of creating language learning activities and general documentation. Most of the language recordings are structured and semi-structured elicitation tasks (translation tasks from English or Dhuwal, picture descriptions, or vernacular descriptions). Five speakers contributed, all of which are now

[5] See https://www.endangeredlanguages.com/revitalization-program/kullilli-ngulkana

[6] As presented in more detail below, we also tried augmenting the Yan-nhangu data with speech samples from another language, Bardi (Nyulnyulan).

[7] See https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface/WebMAUSBasic

Speaker Metadata	Unaugmented File Number	Augmented File Number
Multispeaker	2780	14264
Speaker 3	1691	3011
Speaker 2	418	3658
Speaker 1	266	2321
Speaker 4	0	2358

Table 1. Yan-nhangu audio files by speaker and augmentation

deceased. They represent Warrawarra, Mäl̲arra, and Gamalaŋga clans. Language learning materials include Baymarrwaŋa et al. (2005) and the project produced a draft dictionary for internal community use, which was expanded and published as Baymarrwaŋa & James (2014).

Yan-nhangu recordings were made as translations and semi-structured tasks with five speakers of Yan-nhangu from three of the six Nhangu clans. In the final datasets, which distinguish between multispeaker and individual recordings for *speaker-specific* models, three of these speakers produced recordings by themselves (four in the augmented dataset).

For some models, we augmented the Yan-nhangu data with data from another Australian language, Bardi, in order to test whether multilingual models might be more robust. The Bardi data included four speakers, and a total of 1.38 hours of manually transcribed data.

Manually transcribed audio recordings for Yan-nhangu and Bardi were preprocessed. After removing silences, there are a total of 3.27 hours of human annotated phrase-level transcriptions for Yan-nhangu. An additional 7.19 hours of Yan-nhangu audio were automatically transcribed using a fine-tuned `wav2vec2` model trained on the same human annotated phrase-level transcriptions and included in an augmented dataset to explore the effects of data augmentation (Daul et al. 2026). Automatically generated transcriptions were not manually corrected.[8] For the multilingual acoustic model, 1.38 hours of manually transcribed phrase-level Bardi audio were included in a training dataset. No additional automatically transcribed audio was included for Bardi.

During the collection of the Yan-nhangu and Bardi data, information about the speakers in the recording was also collected. The majority of Yan-nhangu audio contains multiple speakers, while all Bardi audio recordings contain a single speaker. For model training, all Yan-nhangu tracks with more than one speaker are marked as being multispeaker, with no further distinction made with regards to which speakers are included. Table 1 includes information about the amount of data from each speaker for Yan-nhangu, while Table 2 includes information for Bardi.

3.2. CONSENT FOR USE OF MATERIALS. Secondary use of the recordings was discussed as part of informed consent for the project and such questions were regularly revisited while work was underway. As part of the project, community members were clear that it was allowed to use the materials that we had created for other projects, as long as such usage was consistent with a set of general principles that we laid out. This included respect for *Yolŋu Rom* (law), acknowledg-

[8] However, as noted in Daul et al. (2026), such transcriptions usually align well with the human-produced annotations, and where they differ, the differences tend to be either in word divisions or the automatic transcription produces text that is a feasible alternative reading.

Speaker Metadata	File Number
Multispeaker	254
Speaker 4	1302
Speaker 1	374
Speaker 2	340
Speaker 3	223

Table 2. Bardi audio files by speaker

ment of contributions,[9] and obligations to share and elevate Yolngu knowledge while respecting personal information. That is, the original projects included permission for unforeseen secondary uses, as long as they were consistent with cultural principles and brought benefit to Yan-nhangu or other Indigenous communities. A recurrent theme throughout work on the language was reciprocation and an obligation to share Yolngu knowledge. That is, participants made it clear that they were doing this project in part with the expectation that their knowledge and work would be shared widely and appropriately.[10] These views structured most of the activities conducted and all subsequent work with Yan-nhangu materials.

3.3. NON-NEURAL METHODS. In order to test a TTS system, we did two things. We first attempted to create a synthetic voice through eSpeak-NG,[11] a fully synthesized system where a speech grammar of a language is created by specifying mean format measures, segment durations, allophonic rules, and stress rules. In combination with orthographic rules, the resulting grammar is then fed to a synthesizer to produce .wav files. The instructional grammar is produced in text files (see Figure 1 below) and then compiled in C++. Users specify the language, and optionally the voice, and give text that is then synthesized and played or saved as a .wav file.

Such a system has several advantages, along with several substantial disadvantages. A main advantage is the ease with which new languages can be added. The system is fully open source and language rules can be copied and adapted as needed, with language additions being independent of the languages already represented.[12] Once compiled, generating large amounts of audio is very straightforward and it can be incorporated into other scripts and workflows. For example, using eSpeak-NG, one could easily create audio for the headwords and example sentences for a multi-thousand word dictionary in a matter of minutes. Another advantage is that eSpeak voices are not trained on human corpora. That is, it is not necessary to have a large amount of existing speech data (or even *any* data). This is especially important for languages which are known only from written records.

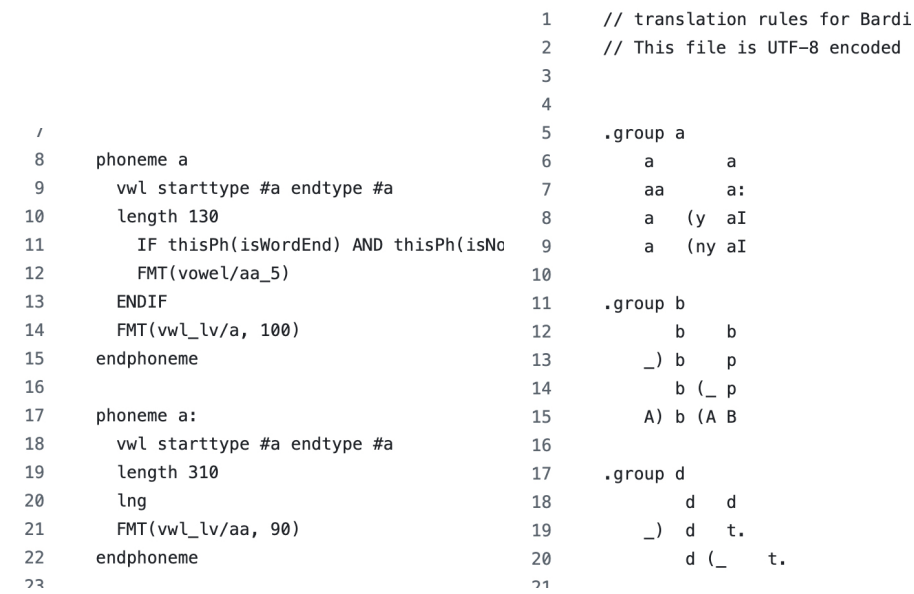
While it is straightforward to create a language, however, it is not straightforward to create naturalistic or even comprehensible speech. In our experiments, for example, it took roughly a day to clone a language, create the orthography files, and compile a "working" voice (in that we could successfully generate wav files from the command line). Several weeks later, we still did

[9] We don't use the names of contributors here because of the Yolngu necronym taboo (prohibition on using the names of people who've passed away).

[10] This came about in part because of an offensive comment by an anthropologist at a Garma Festival in the early 2000s that he thought that Mäl̲arra people were all dead and their language extinct.

[11] https://github.com/espeak-ng/espeak-ng

[12] In many contexts, bootstrapping from closely related languages gives better performance than from unrelated ones (Lam-Yee-Mui et al. 2023; San et al. 2021). This is not the case for eSpeak.



```
phoneme a
  vwl starttype #a endtype #a
  length 130
    IF thisPh(isWordEnd) AND thisPh(isNo
    FMT(vowel/aa_5)
  ENDIF
  FMT(vwl_lv/a, 100)
endphoneme

phoneme a:
  vwl starttype #a endtype #a
  length 310
  lng
  FMT(vwl_lv/aa, 90)
endphoneme
```

```
// translation rules for Bardi
// This file is UTF-8 encoded


.group a
    a       a
    aa      a:
    a   (y  aI
    a   (ny aI

.group b
       b    b
    _) b    p
       b (_ p
    A) b (A B

.group d
        d   d
    _)  d   t.
        d (_    t.
```

Figure 1. Example of eSpeak-NG phonemic instructions. The lefthand side provides instructions for synthesizing /a/ and /a:/. The righthand side gives the allophonic realizations of /a/, /b/ and /d/.

not have a comprehensible voice. Moreover, even when comprehensible, eSpeak voices are not naturalistic (that is, they are obviously synthetic renditions of human speech). We return to this point below.

3.4. NEURAL METHODS. All models were trained using the MatchaTTS architecture (Mehta et al. 2024). To explore the full functionality of the MatchaTTS systems, we trained 8 models, representing every combination of the following parameters.

1. *Monolingual* Yan-nhangu models vs. *Multilingual* Yan-nhangu and Bardi models, containing an additional 1.38 hours of manually transcribed speech.

2. *Non-augmented* models including only the 3.27 hours of manually transcribed Yan-nhangu vs. *Augmented* models including the additional 7.19 hours of automatically transcribed Yan-nhangu (10.46 hours total)

3. *Speaker-specific* models where speaker information is passed in as a parameter during model training vs. *Speaker-general* models where speaker information is not passed in as a parameter. Note that Yan-nhangu has 3 speakers (4 speaker in the *augmented* models' dataset) that contributed individual narratives while Bardi has 4 speakers contributing individual narratives.

Another problem that occurred throughout model training was overfitting. Since it is not trivial to systematically evaluate the quality of generated audio and we were unable to stop the model after a certain amount of epochs, hyperparameter tuning on the number of epochs was performed based on the amount of time the model trained as a proportion of the duration of the training data. We found that training the models for about 50% more time than the duration of the training

data worked best, with all models trained on the same type of GPU to minimize variation. This method is hardware-dependent and a limitation that must be addressed in future iterations.

4. Results.

4.1. ESPEAK-NG. We first produced a "custom" language model using the English settings. This adaptation from an English model produced comprehensible but inaccurate results. That is, entirely unsurprisingly, a TTS model expecting English orthographic text did not produce accurate renditions of another language, but instead produced words pronounced as though they were English words.

Adaptation to Bardi orthography and phoneme durations produced incomprehensible outputs, and output audio representations were not clearly identifiable. That is, such items were not usable for language work. However, we did not come close to exhausting the fine-tuning possibilities for language creation, as we then switched focus to neural methods.

4.2. MATCHA-TTS. Models trained in about 5-8 GPU hours per model. Overall, results of training were excellent and produced clearly interpretable speech. Even the *monolingual non-augmented* models trained on only 3.27 hours of manually annotated Yan-nhangu speech were able to produce a TTS model that generated examples that could immediately be interpreted as Yan-nhangu and transcribed as the intended word.

Monolingual models produced good results for Yan-nhangu. To explore whether we can harness the phonological similarity of Australian languages to in a low-resource language transfer environment, we introduce the *multilingual* models, where every piece of training data is tagged with a language ID and each output must similarly be generated with a language ID. Regardless of whether the models were *augmented* or *non-augmented*, we were unable to produce good results for Bardi. Inclusion of Bardi data additionally did not seem to produce better results for Yan-nhangu. Training *augmented* models seemed to have slightly increased the quality of the generated audio, but impressionistically it was not significantly better.

The biggest improvement in voice quality came from training *speaker-specific* models, where input data was tagged with a unique speaker ID and each output utterance is generated with a speaker ID that needs to be specified. For speakers that contributed greatly to the training data, the result was higher quality speech, though the voices were also much more identifiable. For speakers that contributed less hours of audio, speech was much less intelligible. This created an unfortunate tradeoff, where the best models were the ones that included speaker information, though these same models created much less anonymous speech than the *speaker-general* models. It must be noted that while *speaker-general* models reduced the extent to which the identity of the consultants were identifiable, the composite voice that these models produced had identifiable characteristics from several people. That is, it did not ameliorate the problem, it compounded it.

4.3. ADDITIONAL ANALYSIS. Looking at our three parameters (*monolingual* vs. *multilingual*, *non-augmented* vs. *augmented*, and *speaker-specific* vs. *speaker-general*), we can analyze the way that each parameter influences the outcome.

The parameter with the greatest influence on the quality of the output is *speaker-specific*, with models trained with speaker data outperforming *speaker-general* models. However, this improvement was limited to only the speakers that were most represented in the most training data. Additionally, *speaker-specific* models were worse when producing Bardi speech in the *multi-*

lingual setting. Together, this suggests that training *speaker-specific* models only improves the quality of the outputs for speakers that have large amounts of training data to begin with.

Training *multilingual* models did not seem to improve training accuracy for Yan-nhangu. This may be due to a couple of reasons: (i) Bardi has much less training data than Yan-nhangu, (ii) the Bardi audio quality may be lower than Yan-nhangu's, or (iii) Bardi has many speakers given the amount of audio data available, introducing more variation. Meanwhile, training *augmented* models did seem to result in a slight improvement in the quality of the output, though this was a much smaller difference. The lack of a large improvement may similarly be due to: (i) lower transcription accuracy for the ASR'ed files, (ii) introduction of a new speaker and more multi-speaker tracks. However, even a slight improvement is promising, as it may present an avenue for improving model accuracy in such a low-resource setting.

4.4. INTERIM SUMMARY. The Matcha-TTS models produced clearly interpretable and somewhat naturalistic speech. The voices produced with the speaker-specific models were clearly recognizable as individuals. This was especially true for speakers that contributed most to the datasets.

5. Discussion and implications. As noted in Section 4.2, the voices produced with the speaker-specific models were clearly recognizable as individuals. This was especially true for speakers that contributed most to the datasets. That is, the end result of model training was more like voice-cloning or deep-faking than "text-to-speech" generation. That is, we should think of systems like Matcha-TTS, when used in these settings, as akin to creating synthetic voices of *individuals* rather than simply a creating agnostic multi-modal access to a *language*. Here we explore this topic in more detail.

5.1. THE ISSUE. Generic (non-voice-cloning) TTS treats voices as impersonal or unidentified. Mozilla-TTS, for example, talks about *languages* and *voices*[13] and presents the aim as generating "human-like speech" as contrasted with productions that sound like they were generated by computers. This is perhaps also in the tradition of the use of such systems, where the identity and individual characteristics of cloned voices are not foregrounded.

Indigenous language collections have a very different prioritization of individuality. Contributors are almost always known individuals and are often recording their languages because it's important to them. Indigenous scholars often discuss the importance of treating languages as not simply artefacts divorced from the people who use them, from Perley's (2012) "zombie" language records to Thieberger and Harris' (2022) points about the importance of individuals' legacies in archives.

Because TTS works best with small numbers of voices, naturalistic voices by definition will likely be highly identifiable. We should note further that the Yan-nhangu community members who made recordings of their languages were all well-known in both Milingimbi and Arnhem Land and would probably be recognizable beyond people who know Yan-nhangu. This brings such work into the ethical and philosophical realm of "deepfakes" as well as "speech synthesis" for a *language*. Readers may wonder why we did not anticipate this outcome. In truth, we did not expect the Matcha-TTS models to produce as comprehensible output with as little data as we were working with, and we expected models created from multiple speakers to average out and produce less distinctively individual voices, rather than producing a composite voice with

[13] https://github.com/mozilla/TTS

multiple individual-identifying speech characteristics.

5.2. DID WE DEEPFAKE AN ELDER. In order to evaluate our work with respect to "deepfakes", we use the definition formulated by De Ruiter (2021), which brings together several componentsin the literature. De Ruiter suggests that there are three components that contribution to the creation of a deepfake:

> Three factors are central to determining whether a deepfake is morally problematic: (i) whether the deepfaked person(s) would object to the way in which they are represented; (ii) whether the deepfake deceives viewers; and (iii) the intent with which the deepfake was created. (De Ruiter 2021:1311)

In the computational literature on such work, authors tend to focus on points (ii) and (iii). That is, they situate the ethical issues as ones primarily about lying and deceit (cf. Khanjani et al. 2023; Pawelec 2025). We note, however, that De Ruiter (2021) does not really engage with issues of consent, as objecting to the way someone is represented is not the same as objecting to the synthetic voice being created in the first place.

We suggest that while our voice creations may not technically be deepfakes by De Ruiter's definition, they are nonetheless deeply problematic. We had no intent to deceive listeners; if we had released any of this material, we would have been transparent about the way the material was created. We would have discussed all aspects of the project with language teams and would not have done any further work if the materials had raised concerns. Point (ii) doesn't apply.

Point (iii) does not apply either, since the voices were created with the intent to assist in problem solving and while working in ways that were consistent with previously negotiated consent for secondary use.

We cannot ask those recorded (or their descendants[14]) whether they would object to this work. While we do not speak for them, we suspect these voices would not be viewed as consistent with our agreements or respectful of Yolngu people. We certainly do not view them as consistent with what we agreed to. *Yolngu Rom* includes clear philosophies about the relationship between language, land, and the individual (e.g. Keen 2004). Unidentifiable and obviously synthetic voices (such as the clearly computer-generated voices produced by eSpeak) might be argued to be more clearly outside such relationships, since for such voices there is no individual who learned the language through relationships.

Making voices that might be associated with individuals also ignores the sociolinguistics of word and register choice in Yolngu linguistics. For example, Yolngu Matha language ideologies include *bokmakku matha* — words that are common to many Yolngu clans and do not convey any particular clan association — but also words that are *l̪imuruŋ* "ours". These are clan-specific words, usage of which by members of other clans would be considered appropriative (or a violation of social norms). There are also individual-specific taboos, such as necronym taboos (the prohibition on saying the names of deceased relatives). For example, recording dictionary headwords or examples at scale would very likely have the synthetic voice saying words that the real person would not say.

We believe that creating such voices would be a violation of traditional Yolngu views of respectful behavior towards deceased people and their legacies, at least according to how those

[14] We have tried to contact family members but have been unable to do so.

views were explained.We also feel that it's disrespectful within our own ethical and moral codes. As described above, the negotiation of consent to do language work included an expectation of knowledge sharing (that Yolngu knowledge would be discussed in university classes and attributed, for example). As part of learning about appropriate custodianship of language materials and how to share knowledge appropriately, points about how to treat these legacies were explicitly discussed. It included permission to play audio recordings in respectful educational contexts, but it did not include, for example, showing photographs.[15]

While our discussions around consent did include various contexts for secondary use, the types of secondary uses that we envisaged were very different from the creation of synthetic identifiable voices. The discussions focused on data, tools, and experience, not identities. That is, they included points like whether it was all right to show the learner's guide we'd created (Baymarrwaŋa & James 2014) as an example to other Indigenous communities who were thinking about language work, or whether it was all right to use analytical tools such as formant extraction scripts on multiple languages.

In summary, good intentions and lack of intent to deceive do not cancel out the other substantial problems created by such voices, whether or not such voices were created inadvertently.

6. Conclusions and Recommendations.

6.1. SUMMARY. TTS development has come a long way, and it's now possible to create naturalistic voices with only a few hours of data and with intermediate technical skills. This makes them tempting for language reclamation and support projects, particularly where speechification of print-only data would be useful. Such voices, however, are not simply "naturalistic", they are *identifiable*, making them more akin to deepfakes or voice cloning than speech synthesis. We argue that good underlying motivations and lack of intent to deceive do not cancel out the negatives of creating these synthetic voices. This situation has arisen, in part, because TTS is framed as converting between language modalities, rather than creating individual voices. It is particularly acute when the language materials come from archives and were recorded with people who are now deceased.

6.2. RECOMMENDATIONS. We close with some recommendations. Given the rapidly expanding field of AI avatars and deadbots/thanatobots, it's urgent to have clear acceptable usage rules in place about acceptable use of archival materials. Many archives have usage guidelines which include no manipulation of audio, but it's not clear that training data of this type would fall under those definitions of manipulation of the original data, since neural-based models do not directly touch the original recordings.

For researchers and community members, if you wish to create avatars:

- consult explicitly about whether it's a good idea, and specifically consider the implications of doing so, given the sociolinguistic norms and ideologies in use in the community;
- get explicit consent and create custom datasets with full participation of relevant individuals and communities;

[15] There is no information, to our knowledge, on Yolngu responses to hearing synthetically generated voices or viewing AI generated avatars of those who have passed away, known as thanatobots or deadbots. However work in other cultures commonly describes such items using adjectives such as awful, horrible confronting, distressing, and creepy (cf. Hesse 2025).

- include examples of the end product as part of the consent process;
- have clear constraints on what such avatars can be used for and who can use them, and revisit this regularly;
- make it clear on all materials created with the voice or avatar that it is AI-based.
- Get explicit permission (or denial of permission) on archival collections from those close to the individuals in the recordings, if not the individuals themselves.

If you can't get this level of consent and interaction and still need TTS, for example, because of lack of access to the relevant individuals or because the project involves a cross-language model, use eSpeak or another fully synthetic method.

We also strongly recommend that language archives consider making AI provisions a distinct part of consent to use collections, even for otherwise "public" materials.

References

Aklif, Gedda. 1999. *Ardiyooloon Bardi ngaanka: One Arm Point Bardi dictionary*. Halls Creek, Western Australia: Kimberley Language Resource Centre.

Baymarrwaŋa, Laurie, Rita Gularrbanga, Laurie Milinditj, Rayba Nyaŋbal, Margaret Nyuŋunyuŋu, Allison Warrŋawun & Claire Bowern. 2005. *A learner's guide to Yan-nhaŋu*. Milingimbi: Milingimbi Literature Production Centre.

Baymarrwaŋa, Laurie & Bentley James. 2014. *Yan-nhaŋu atlas and illustrated dictionary of the Crocodile Islands*. The Tien-Wah Press, Singapore & Sydney Australia.

Bowern, Claire (ed.). 2023. *The Oxford Guide to Australian languages*. Oxford University Press.

Bowern, Claire, Linda Lanz & David Katten (eds.). 2007. *A learner's guide to Bardi*. IAD Press.

Daul, Massimo Marie, Alessio Tosolini & Claire Bowern. 2026. Linguistically informed tokenization improves ASR for underresourced languages. In Éric Le Ferrand, Elena Klyachko, Shu Okabe, Ekaterina Voloshina, Oleg Serikov, Tatiana Shavrina & Ekaterina Vylomova (eds.), *Proceedings of the fifth workshop on NLP applications to field linguistics*, 31–37. Rabat, Morocco: Association for Computational Linguistics. https://doi.org/10.18653/v1/2026.fieldmatters-1.4. https://aclanthology.org/2026.fieldmatters-1.4/.

De Ruiter, Adrienne. 2021. The Distinct Wrong of Deepfakes. *Philosophy & Technology* 34(4). 1311–1332. https://doi.org/10.1007/s13347-021-00459-2. https://link.springer.com/10.1007/s13347-021-00459-2.

Gölge, Eren & Coqui AI Development Team. 2025. Coqui TTS. https://coqui.ai/.

Gumma, Varun, Rishav Hada, Aditya Yadavalli, Pamir Gogoi, Ishani Mondal, Vivek Seshadri & Kalika Bali. 2024. MunTTS: A Text-to-Speech System for Mundari. In Sarah Moeller, Godfred Agyapong, Antti Arppe, Aditi Chaudhary, Shruti Rijhwani, Christopher Cox, Ryan Henke, Alexis Palmer, Daisy Rosenblum & Lane Schwartz (eds.), *Proceedings of the Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages*, 76–82. St. Julians, Malta: Association for Computational Linguistics. https://aclanthology.org/2024.computel-1.11/.

Hammerly, Christopher, Sonja Fougre, Giancarlo Sierra, Scott Parkhill, Harrison Porteous & Chad Quinn. 2023. A text-to-speech synthesis system for Border Lakes Ojibwe. In Atticus Harrigan, Aditi Chaudhary, Shruti Rijhwani, Sarah Moeller, Antti Arppe, Alexis Palmer, Ryan Henke & Daisy Rosenblum (eds.), *Proceedings of the Sixth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, 60–65. Remote: Association for Computational Linguistics. https://aclanthology.org/2023.computel-1.9/.

Harrigan, Atticus, Timothy Mills & Antti Arppe. 2019. A Preliminary Plains Cree Speech Synthesizer. *Proceedings of the Workshop on Computational Methods for Endangered Languages* 1. https://doi.org/10.33011/computel.v1i.421. https://journals.colorado.edu/index.php/computel/article/view/421.

Hesse, Monica. 2025. Column | I used an AI tool to resurrect my grandmother, and it was awful. *The Washington Post* https://www.washingtonpost.com/style/2025/10/13/ai-loved-ones-grandmother/.

Holmer, N.M. 1988. *Notes on some Queensland languages*, vol. D79. Canberra: Pacific Linguistics.

Keen, Ian. 2004. Aboriginal economy and society: Australia at the threshold of colonisation.

Khanjani, Zahra, Gabrielle Watson & Vandana P. Janeja. 2023. Audio deepfakes: A survey. *Frontiers in Big Data* 5. 1001063. https://doi.org/10.3389/fdata.2022.1001063. https://www.frontiersin.org/articles/10.3389/fdata.2022.1001063/full.

Kim, Jaehyeon, Jungil Kong & Juhee Son. 2021. Conditional Variational Autoencoder with Adversarial Learning for End-to-End Text-to-Speech. https://doi.org/10.48550/arXiv.2106.06103. ArXiv:2106.06103 [cs]. http://arxiv.org/abs/2106.06103.

Lam-Yee-Mui, La-Marie, Waad Ben Kheder, Viet-Bac Le, Claude Barras & Jean-Luc Gauvain. 2023. Multilingual Models with Language Embeddings for Low-resource Speech Recognition. In *Proceedings 2nd Annual Meeting of the ELRA/ISCA SIG on Under-resourced Languages (SIGUL 2023)*, 83–87. Dublin, Ireland: ISCA. https://doi.org/10.21437/SIGUL.2023-18. https://universite-paris-saclay.hal.science/hal-04397807. Backup Publisher: ELRA/ISCA.

McDonald, Maryalice & Stefan Wurm. 1979. *Basic materials in Wangkumara (Gaḷali)*, vol. B-65. Canberra: Pacific Linguistics.

Mehta, Shivam, Ruibo Tu, Jonas Beskow, va Szkely & Gustav Eje Henter. 2024. Matcha-TTS: A fast TTS architecture with conditional flow matching. https://doi.org/10.48550/arXiv.2309.03199. ArXiv:2309.03199 [eess]. http://arxiv.org/abs/2309.03199.

Mengke, Dalai, Yan Meng & Péter Mihajlik. 2026. How far can synthetic speech go? enhancing asr inlow-resource scenarios viavoice cloning. In Kamil Ekštein, Miloslav Konopík, Ondřej Pražák & František Pártl (eds.), *Text, speech, and dialogue*, 207–217. Cham: Springer Nature Switzerland.

Oord, Aaron van den, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior & Koray Kavukcuoglu. 2016. WaveNet: A Generative Model for Raw Audio. https://doi.org/10.48550/arXiv.1609.03499. ArXiv:1609.03499 [cs]. http://arxiv.org/abs/1609.03499.

Pawelec, Maria. 2025. Decent deepfakes? Professional deepfake developers ethical considerations and their governance potential. *AI and Ethics* 5(3). 2641–2666. https://doi.org/10.1007/s43681-024-00542-2. https://link.springer.com/10.1007/s43681-024-00542-2.

Perley, Bernard C. 2012. Zombie Linguistics: Experts, Endangered Languages and the Curse of Undead Voices. *Anthropological Forum* 22(2). 133–149. https://doi.org/10.1080/00664677.2012.694170. http://www.tandfonline.com/doi/abs/10.1080/00664677.2012.694170.

Pine, Aidan, Erica Cooper, David Guzmn, Eric Joanis, Anna Kazantseva, Ross Krekoski, Roland Kuhn, Samuel Larkin, Patrick Littell, Delaney Lothian, Akwiratkha Martin, Korin Richmond, Marc Tessier, Cassia Valentini-Botinhao, Dan Wells & Junichi Yamagishi. 2025. Speech Generation for Indigenous Language Education. *Computer Speech & Language* 90. 101723. https://doi.org/10.1016/j.csl.2024.101723. https://www.sciencedirect.com/science/article/pii/S0885230824001062.

Pine, Aidan, Dan Wells, Nathan Brinklow, Patrick Littell & Korin Richmond. 2022. Requirements and Motivations of Low-Resource Speech Synthesis for Language Revitalization. In Smaranda Muresan, Preslav Nakov & Aline Villavicencio (eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7346–7359. Dublin, Ireland: Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.acl-long.507. https://aclanthology.org/2022.acl-long.507/.

Ren, Yi, Chenxu Hu, Xu Tan, Tao Qin, Sheng Zhao, Zhou Zhao & Tie-Yan Liu. 2022. FastSpeech 2: Fast and High-Quality End-to-End Text to Speech. https://doi.org/10.48550/arXiv.2006.04558. ArXiv:2006.04558 [eess]. http://arxiv.org/abs/2006.04558.

Round, Erich R. 2023. Segment inventories. In Claire Bowern (ed.), *The Oxford Guide to Australian Languages* Oxford Guides to the World's Languages, Chapter 10. Oxford, New York: Oxford University Press.

San, Nay, Martijn Bartelds, Mitchell Browne, Lily Clifford, Fiona Gibson, John Mansfield, David Nash, Jane Simpson, Myfany Turpin, Maria Vollmer, Sasha Wilmoth & Dan Jurafsky. 2021. Leveraging Pre-Trained Representations to Improve Access to Untranscribed Speech from Endangered Languages. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 1094–1101. https://doi.org/10.1109/ASRU51503.2021.9688301. https://ieeexplore.ieee.org/document/9688301.

Shen, Jonathan, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, R. J. Skerry-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis & Yonghui Wu. 2018. Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions. https://doi.org/10.48550/arXiv.1712.05884. ArXiv:1712.05884 [cs]. http://arxiv.org/abs/1712.05884.

Thai, Bao, Robert Jimerson, Dominic Arcoraci, Emily Prud'hommeaux & Raymond Ptucha. 2019. Synthetic data augmentation for improving low-resource asr. In *2019 ieee western new york image and signal processing workshop (wnyispw)*, 1–9. https://doi.org/10.1109/WNYIPW.2019.8923082.

Thieberger, Nick & Amanda Harris. 2022. When Your Data is My Grandparents Singing. Digitisation and Access for Cultural Records, the Pacific and Regional Archive for Dig-

ital Sources in Endangered Cultures (PARADISEC). *Data Science Journal* 21. 9. https://doi.org/10.5334/dsj-2022-009. http://datascience.codata.org/articles/10.5334/dsj-2022-009/.

Wang, Shenran, Changbing Yang, Michael l Parkhill, Chad Quinn, Christopher Hammerly & Jian Zhu. 2025. Developing multilingual speech synthesis system for Ojibwe, Mi'kmaq, and Maliseet. In Luis Chiruzzo, Alan Ritter & Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, 817–826. Albuquerque, New Mexico: Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.naacl-short.69. https://aclanthology.org/2025.naacl-short.69/.

Wang, Yuxuan, R. J. Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J. Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, Quoc Le, Yannis Agiomyrgiannakis, Rob Clark & Rif A. Saurous. 2017. Tacotron: Towards End-to-End Speech Synthesis. https://doi.org/10.48550/arXiv.1703.10135. ArXiv:1703.10135 [cs]. http://arxiv.org/abs/1703.10135.

Whitman, Robert, Richard Sproat & Chilin Shih. 1997. A Navajo Language Text-to-Speech Synthesizer. Tech. rep. AT&T Bell Labs.