UNIVERSITY OF ALBERTA

Malagasy Speech Synthesis

BY

Tyler T. Schnoor

A THESIS

SUBMITTED TO THE FACULTY OF ARTS

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

BACHELOR OF ARTS

DEPARTMENT OF LINGUISTICS

EDMONTON, ALBERTA

April, 2022

UNIVERSITY OF ALBERTA

FACULTY OF ARTS

The undersigned certify that they have read and recommend to the Faculty of Arts for acceptance, a thesis entitled Malagasy Speech Synthesis, submitted by Tyler T. Schnoor in partial fulfillment of the requirements for the degree of Bachelor of Arts.

…….

Honors Thesis Supervisor

…….

For the Department

ABSTRACT

Speech technologies may benefit people by improving the accessibility of information or services, by increasing productivity, or by generally improving human-computer interaction. However, speech technologies are only available for use in a small portion of the world's languages. The present study aims to investigate some of the means by which contemporary machine learning approaches to speech synthesis may be adapted for use with under-resourced languages which do not have abundant data available. The first objective of the present study is to develop a Malagasy speech synthesis model which is effective enough to have practical implications. The second research objective is to explore whether the addition of crowd-sourced training data is beneficial to the model. I develop a web application which facilitates the remote collection of speech data and use it to collect a small, multi-speaker, Malagasy speech dataset for use in training. The merits of crowd-sourcing data from multiple speakers are compared to the merits of collecting data from a single speaker. The third and final objective is to explore the effects of cross-lingual transfer learning, data augmentation, and other methods which might facilitate the development of speech synthesis models for under-resourced languages. This is done by iteratively training models using these methods and comparing their outputs. The models are made using an open source implementation of the Tacotron framework (*Tacotron 2 (without Wavenet)*, 2018/2022). The results of the present study suggest that a combination of cross-lingual transfer learning and data augmentation methods may be employed to train effective speech synthesis models using a small amount of speech data in an under-resourced language. The addition of multi-speaker data is not found to improve results when combined with a small single-speaker training set. Further investigation will determine whether multi-speaker data may be incorporated in other ways to enhance model outputs.

ACKNOWLEDGEMENTS

I would like to express my gratitude to my honors supervisor, Dr. Benjamin V. Tucker for all of his support and guidance throughout this project and Dr. Jorges Rosés-Labrada for feedback on the paper as the second reader. I would also like to thank the University of Alberta Undergraduate Research Council for funding the data collection portion of this project through the Roger S. Smith Undergraduate Student Researcher Award. I am grateful to Tafitasoa Rasolofonjatovomalala for her help in translating. Additionally, I am grateful to the members of the Alberta Phonetics Laboratory and any others who gave feedback on the project. I gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan V GPU used for this research.

TABLE OF CONTENTS

CHAPTER I. Introduction

1.0 Malagasy

Malagasy is the native language of the inhabitants of Madagascar who are descended from the people of Borneo (Brucato et al., 2016). There are currently approximately 25 million native Malagasy speakers. The Malagasy language is most closely related to Maanjan and other Bornean languages, and has been influenced by Sanscrit and—more significantly—Bantu. Although researchers are studying the language (Adelaar, 2018; Aziz, 2020; Aziz & Paul, 2019; Potsdam, 2022; Serva & Pasquini, 2021), those areas of research which require large datasets (e.g. language technology, corpus linguistics) will be difficult to pursue until such datasets become more readily available. Although recent work has been published in the realm of Malagasy text summarization with deep learning (Ratianantitra et al., 2019), there are no Malagasy language technologies available for use by the average Malagasy person at the time of writing. The goal of the present study is to investigate methods of developing text-to-speech (TTS) models for under-resourced languages. This is done by iteratively training TTS models using techniques for maximizing the usefulness of a small Malagasy speech dataset.

Malagasy is a good candidate for the development of speech technologies, especially in the context of development for under-resourced languages. Although there are relatively limited resources available for the task of developing speech technologies for Malagasy, there are many native speakers who might benefit from these technologies being available in their language. This is representative of the situation of many language communities worldwide. Glottolog notes that 25.6% of languages in its database have a Most Extensive Description of “Wordlists or less”

(Hammarström et al., 2021), showing that many languages are under-resourced purely by descriptive standards. Although I do not know of any comprehensive study into the number of languages for which openly accessible speech corpora have been created, my impression is that the number is low. This number is even more limited when considering that few academic corpora are collected with specific ethics approval for use in the development of speech technologies. It is therefore valuable to study the measures by which a limited amount of speech data may be optimized for use in training machine learning models.

1.1 Research Objectives

The first part of this project is focused on collecting a Malagasy speech dataset that will enable further research on the Malagasy language. There is little open-access data on the language as it is, so a corpus of short utterances is a valuable resource for many forms of research. As the end goal of my honors thesis is to create a speech synthesis model, the dataset is formatted to suit speech technology applications. However, the data also includes the necessary information to make it useful for phoneticians, descriptive linguists, and other language researchers.

Conditions created by the Covid-19 pandemic have severely affected many in-person research activities (Tuttle, 2020). Speech data collection is certainly one aspect of linguistic research which is made more difficult by social distancing measures. Even in times when social distancing is unnecessary, many factors like geographic distance from a population of interest and lack of equipment may make traditional data collection methods difficult. A less used but increasingly viable option for certain data collection needs may be found in remote methods of

speech data collection, or crowd-sourcing. Increasing internet access (Ronquillo & Currie, 2012) and improving internet browser functionality continue to make online data collection more feasible for various types of research. New mobile devices are compatible with modern browsers and have high-quality, integrated microphones, meaning the required tools for volunteering speech data are already in the pockets of many. For these reasons, I developed a web application which facilitates the remote collection of speech data for researchers. I describe the application further in section 2.1.

There are three main objectives for the second part of this project, which is focused on the exploration of strategies for developing deep neural network models of speech synthesis for under-resourced languages:

1. Apply an existing end-to-end text-to-speech deep neural network framework (Tacotron [Wang et al., 2017]) to Malagasy in order to synthesize speech in a practical way
2. Explore whether the addition of crowd-sourced training data is beneficial to the model
3. Explore the effects of cross-lingual transfer learning, data augmentation, and other methods which might facilitate the development of speech synthesis models for under-resourced languages

One popular and effective method by which a model may be trained to synthesize speech in an under-resourced language is to build on text-speech associations which have already been learned from training on a different language. By doing so, a model can use one language's sounds as a starting point rather than starting with random values. These practices which involve multiple languages in the training of models are known as multilingual data strategies. Do et al. (2021) offer a systematic review of text-to-speech models which are trained using multilingual

data strategies. In the case of machine learning, this takes the form of cross-lingual transfer learning. Cross-lingual transfer learning is an example of a multilingual data strategy in which a neural network is first trained using source data from a source language (usually a language with abundant data), and is subsequently fine-tuned using target data from a target language (an under-resourced language, for example). Because there is always overlap in some text-speech associations between languages, the source language may fill informational voids which would otherwise be caused by data sparsity. Ideally, the result of cross-lingual transfer learning is a model which is made more robust by the source data but which synthesizes speech that matches the speaker characteristics and sound inventory of the target data.

Do et al. gather many instances of cross-lingual transfer learning in the development of TTS models for under-resourced languages and endeavor to provide a method by which one might compare their results. The present study is to be another example of multilingual strategies while also addressing the interplay of data characteristics (e.g. whether data are augmented, whether data are from a single speaker) and the practicality of certain strategies for engineers who may not have domain expertise in linguistics.

1.2 Motivations

Speech technologies—principally speech synthesis and automatic speech recognition—can offer a number of benefits if they are available in one's language. These benefits may take the form of tools which help increase productivity or efficiency, like a TTS model which can be used to generate automated announcements over a speaker system, for example. There are also applications in human-computer interaction more generally. Some enjoy

the convenience of being able to interact with a device from a distance or while doing other tasks, like cooking. Recent advances in mobile computing and the Internet of Things (Li et al., 2015) have put digital voice assistants—whose core functionality relies on algorithms for speech synthesis and recognition—into cars, speakers, and smartphones (Polyakov et al., 2018).

Speech technologies have also transformed businesses. Call centres and remote customer services are using speech synthesis and recognition algorithms in combination with conversational artificial intelligence to automate their businesses (Qiu & Benbasat, 2005). The improvement of synthesized speech quality has also allowed for the use of TTS models in video games (Farner et al., 2008) and other forms of entertainment.

One of the most transformative applications for speech technologies is in improving accessibility (Freitas, 2010). Some of those most benefited by the increase in accessibility are those with special needs. Augmentative or alternative communication devices, for example, use speech synthesis to offer a method of communication to those who may be unable to communicate otherwise (Higginbotham, 2010; Mills et al., 2014; Řepová et al., 2021). There has also been an effort to improve accessibility in the production of entertainment and services. For example, streaming and broadcasting services have incorporated automatic transcription as a cost-effective way of making media accessible to those with visual impairments (S. S. Chen et al., 2002; Liao et al., 2013).

Although speech technologies are becoming common and readily available for speakers of English or other major languages, the majority of the world's languages have little to no speech technology resources available (Anumanchipalli & Black, 2010; Besacier et al., 2014). This is mainly due to the relative lack of publicly available, properly formatted datasets. Speech

technology companies often pass up opportunities to gather resources for lesser-known languages such as Malagasy in favour of more widespread, lucrative markets. The development of effective speech technologies using limited data often requires special consideration and different methods compared to situations where data are abundant. Studying these methods and disseminating findings may be key to making speech technologies available in more languages.

CHAPTER II. Data Collection

This chapter includes a description of the collection and composition of datasets which are generally used for training TTS models as well as a description of the data which are used in the present study. I begin with an overview of applications—native and web-based—which have been used to collect speech corpora, their strengths and weaknesses, and explain why I chose to create my own web application to facilitate the crowd-sourcing of speech data. I then address the advantages and disadvantages of crowd-sourced data before describing the recordings which were crowd-sourced for the present study. The chapter concludes with a discussion of the advantages and disadvantages of single-speaker data and a description of the single-speaker data used in the present study.

2.0 Data Collection Applications

There have been successful efforts to make native applications for speech data collection in the last decade (Blachon et al., 2016; de Vries et al., 2014). The term native application is used in the present study to refer to an application which runs natively on a certain type of device. Typically, native applications are downloaded from a webpage or an app store onto a smartphone or tablet. Because native applications run natively on one's own device, they do not necessarily require an internet connection to function.

One example of a native application for speech data collection is the Android application called Woefzela, which was used to collect over 800 hours of speech in eleven South African languages for use in automatic speech recognition datasets (de Vries et al., 2014). Similarly, the

Aikuma mobile application was used to collect over 80 hours of speech data in the Congo-Brazzaville area, mainly with a language documentation focus (Blachon et al., 2016). This method of developing an application which is native to certain mobile devices has proven effective, especially when researchers are physically present to facilitate the data collection process and demonstrate to participants how the application is used. These applications can take advantage of features which are native to the device or may be able to utilize more of the device's processing power. The fact that this method allows for data collection to proceed even when no internet connection is available is particularly advantageous in developing regions where internet access may not be widespread. The function of speech data collection applications is first to elicit an utterance. Typically, this is done with a text prompt. However, Aikuma features methods of eliciting utterances for unwritten languages (e.g. by showing picture prompts on the screen). The application records the spoken utterance and the new data must then be saved and organized in such a way that it can be used for research later on.

A major disadvantage associated with the native application approach is that the application can only be used on a device for which it was made. This is not a problem if the researcher wishes to record speech data using their own device and is able to transport the device to each participant. If recordings must be made remotely, a device cannot be transported to the participants, and participants must use their own devices to make recordings; however, compatibility issues between devices may quickly limit the participant pool from which data can be drawn. A web application, on the other hand, is run entirely in a user's browser and is accessed in the same way one accesses other webpages. This alleviates many concerns regarding compatibility, as most browsers work with most devices. Additionally, this means participants can use desktop computers and laptops to access the application as well as mobile devices. Web

applications generally require minimal changes in order to be compatible with various browsers and devices, but native applications may need to be entirely or partially rewritten in a different programming language. For these reasons, I believe the web application approach is the best way to collect data remotely from a wide range of devices in a cost effective manner.

Web applications have already proven successful in the remote collection of speech data. One of the most successful projects is VoxForge, a platform designed to collect speech for open-source speech recognition engines (*VoxForge*, 2006). Another application, called Voicer, was developed with a very similar end goal as the data collection efforts in the current study—to gather speech corpora for under-resourced languages through a web application (Buddhika et al., 2018). Voicer was used to collect 10 hours of Sinhala speech in the form of 39 sentences commonly used in banking transactions. Other projects have outlined the potential for integrating games into speech data collection in order to make participation feel more rewarding (Dumitrescu et al., 2014).

Although the applications above are examples of successful implementations of the web application approach, I could not use them for data collection in the present study for three reasons. First, although VoxForge is open-source, it is somewhat outdated and contains functionalities which are outside the scope of the present study. These functionalities might present issues for the programmer who would modify a version of VoxForge for a different project, as modification to one part of the code might result in unexpected errors elsewhere. Second, as far as I can tell, Voicer is not available in a public repository or hosted in a way that makes it available for use by other researchers. Finally, none of the web applications I found were designed to accommodate additional, self-contained data collection efforts beyond their original use.

2.1 SpeakerPool

The web application I developed, called SpeakerPool, is designed to meet the needs of the present study which are not met by existing applications. SpeakerPool provides some of the same functions as existing methods: a recording interface is provided, prompts and participants are assigned identification numbers, and prompts are shown on the screen for the participant to read aloud. However, there are some key differences which make SpeakerPool a better option for the crowd-sourcing of speech data. First, SpeakerPool is a web application, and therefore runs entirely in the participant's browser, regardless of the device type or operating system. As pointed out in section 2.0, this means that there are less compatibility limitations. An additional advantage of the web application format is that the data is automatically uploaded to a secure server immediately after being recorded, so data is less likely to be lost. This is an advantage over some native applications which require data to be transferred manually. The web application provides other features that are useful to researchers, many of which are not available in the existing solutions, such as a secure login system, unique participant ID assignment, stimuli enumeration, microphone logging, consent form embedding, and the capability to switch all instructions, buttons, and other text to a different language.

My approach is similar to that used in developing Voicer (Buddhika et al., 2018). However, one of the primary goals I had for my application was to maximize reusability. Should I (or another researcher) want to add another study to the application and receive ethics approval to do so, I need to be able to run each with its own consent form, demographic survey, and prompt list. To my knowledge, the necessary functionalities which would allow for the

simultaneous hosting of multiple self-contained studies are not built into any existing data collection solutions.

Crowd-sourcing in the context of data collection is a process by which a large number of volunteers contribute a small amount of data in order to create a dataset (Parent & Eskenazi, 2011). SpeakerPool was designed to facilitate the crowd-sourcing process by making the opportunity to contribute data available to as many people as possible. This is done by making SpeakerPool available for use at any time, accessible via the internet from anywhere, and by minimizing device compatibility issues. Participants are also able to contribute as much or as little data as they would like and are able to stop and restart recording as needed. This flexibility is implemented so that even those participants who only have a short amount of time to record themselves are able to contribute their data.

2.2 Crowd-Sourced Data

The crowd-sourcing approach is advantageous in that a researcher can collect a large amount of data with relative ease, especially if a platform like SpeakerPool is already available for use. A core part of applications such as SpeakerPool is that their functionality is automated, and therefore available for use at any time via the internet. This is very different from traditional methods of collecting speech data, which often require an investigator present to operate recording equipment and to give directions. Additionally, crowd-sourcing the data puts less of a burden on each individual participant. By collecting a small amount of data from each participant (or by allowing each participant to choose how much data they contribute), a researcher is able to

gather a large corpus without demanding a significant contribution of time or effort from any single participant.

Although crowd-sourcing does away with some of the demands placed on investigators and participants, it also means the quality of the data is less controlled. Being a web application, SpeakerPool is available for use in virtually any environment in which a device can access the internet, regardless of levels of background noise, microphone quality, acoustic properties of the room, etc. Thus, the crowd-sourcing approach may yield a larger dataset than traditional methods, but the data will feature more variation and noise than traditional methods.

Researchers may do a number of things in order to encourage participants to improve the quality of their recordings. For example, the ideal recording environment might be described in the instructions read by participants before using the application. They might be encouraged to record themselves when alone in a quiet environment indoors, to keep the microphone a consistent distance from their mouths, and try to maintain a normal pace when reading. The inclusion of certain features in the application can reduce the amount of low quality data recorded in the first place. One such feature in SpeakerPool is an embedded audio player which allows participants to play back their recordings in order to ensure good quality. Measures may also be taken so that low-quality data can be easily removed later on. For example, microphone names are logged by SpeakerPool so that the data recorded on low-quality or malfunctioning microphones can be identified and removed easily.

Crowd-sourced speech data is only suitable for use in certain kinds of research and in the development of certain technologies. For example, most of the referenced projects involving crowd-sourced speech data were started with automatic speech recognition (ASR) in mind.

Crowd-sourced data is particularly well-suited for use with ASR models because it represents a wide selection of the various speaker characteristics and speech environments which the model will encounter. In fact, especially when it comes to developing machine learning models, presenting the model with a diverse dataset at the time of training is a key part of avoiding bias from manifesting in its use case (Mehrabi et al., 2021). However, the high amount of variance in crowd-sourced data makes it less desirable for use in the development of speech synthesis or text-to-speech (TTS) models. Although there is some level of consistency across speakers of the same language, speaker characteristics like speech rate, pitch range, intonation, and the realization of certain speech sounds vary greatly. This presents a significant challenge to the machine learning engineer whose aim is to create a model which can consistently synthesize a single speaker's style of speech. If trained on a combination of different speech styles, the model is given an "unsolvable problem". In such cases, the model learns an intermediate style of speech (an average between the various styles presented at training) which may sound unnatural.

One of the goals of this project is to determine whether there are methods by which one might overcome the aforementioned challenges associated with training machine learning speech synthesis models with small, highly variable datasets. If it were possible to use a crowd-sourced speech dataset in the training of both TTS and ASR models, the investment of gathering such data would grow far more valuable. It might make efforts to gather data more attractive to potential funding agencies if it is known that both technologies may be developed as a result. I hypothesize that crowd-sourced data will not be usable in the development of a TTS model without either heavy manual editing to normalize the data or used in conjunction with a single-speaker dataset.

An overview of the crowd-sourced data used in the current study may be found in Table 1. The dataset amounts to 378 recordings from five participants, two male and three female. The cumulative duration of the data is 0.78 hours. All participants are native Malagasy speakers between 20 and 25 years of age. With the exception of one who recorded at a sampling rate of 44100 Hz and a bitrate of 705 kbps, all participants recorded at a sampling rate of 48000 Hz and a bit rate of 768 kbps. This inconsistency in recording parameters is a result of participants using various devices and microphones to contribute data. All recordings are saved in the .wav file format. The lowest number of recordings contributed by a single participant is 29 and the highest is 102. The participants were asked to record themselves reading aloud the prompt given on their screen. The prompts consisted of random sentences from Malagasy Wikipedia pages (Wikipedia, 2022). Participants were allowed to skip any prompt.

Participant	**Sex**	**Age**	**Sampling Rate (Hz)**	**Bit Rate (kbps)**	**Recordings Contributed**
1	F	20	48000	768	29
2	F	25	48000	768	99
3	M	20	48000	768	102
4	F	20	44100	705	86
5	M	20	48000	768	62

Table 1: Participant numbers are assigned to all participants in the crowd-sourced dataset and are shown in the first column of the table. The reported sex and age of each participant are listed in the second and third columns. The sampling rate and bit rate of the participants' recordings are presented in the fourth and fifth columns. The number of recordings contributed by each of the participants are listed in column six.

Although the description of the dataset is accurate at the time of writing, I plan to continue to recruit participants indefinitely. I will also continue to update and improve the SpeakerPool platform so that it is more useful and accessible to researchers and participants alike.

2.3 Single-Speaker Data

In many ways, gathering a dataset from a single speaker has the opposite advantages and disadvantages when compared to the crowd-sourcing approach. A participant who is tasked with providing a dataset consisting of only their own speech faces a much higher time demand. Because of this, it may be difficult to find participants who are willing to make such a contribution, even if they are compensated for their time. However, the end product is a consistent and controlled dataset which might be used for many types of research, but is especially conducive to training machine learning models for speech synthesis.

There is nothing preventing the use of a web application such as SpeakerPool from being used to collect a substantial amount of data from a single speaker, and it may be reasonable to do so if all speakers of the language of interest are geographically separated from the investigator. However, the consistently high quality of single-speaker data may be lost if an investigator is not present to give direction and run the recording equipment. This is why single-speaker recordings are most commonly collected in a traditional recording booth environment. Alternatively, an investigator might draw from existing materials in order to build a dataset. Audiobooks are used for training machine learning models because they are often recorded using high quality equipment and in a consistent recording environment (Panayotov et al., 2015). The

single-speaker data used in the present study are drawn from Malagasy audiobook recordings for these reasons.

The single-speaker dataset used in the present study was sourced from a website called nybaiboly.net (*Ny Baiboly Malagasy*, n.d.) which functions as a repository for free-to-use Bible media in both French and Malagasy. Among the available resources are audiobook recordings of a single Malagasy speaker reading The Bible in its entirety. It is this audiobook recording which I decided to use for the present study. In order to split the recordings into sections of the size expected by the TTS model, I manually cut them into snippets which are approximately 6-12 seconds long. Attention was paid to making recording boundaries at punctuation marks or natural pauses and the corresponding text for each recording was organized in a document. 284 utterances amounting to 0.58 hours in total duration were separated in this way, all taken from the first seven chapters of Matthew. 243 recordings were made into a training set (the first six chapters) and 41 were made into a validation set (the seventh chapter).

CHAPTER III. Speech Synthesis

This chapter describes the process of training a Malagasy TTS model using the data discussed in chapter two. I begin with a brief introduction to TTS generally, followed by a description of the Tacotron framework (Wang et al., 2017) which is used in the present study. Next, I discuss the methods which are implemented in order to improve model outputs and evaluate the corresponding models. The chapter concludes with a general discussion of the results.

3.0 Text-to-Speech and the Tacotron Framework

In the present study, speech synthesis and text-to-speech models may be thought of as performing the same function: the synthesis of speech from text inputs. TTS models have been an area of intense research which has given rise to a number of methods by which speech synthesis can be performed (Tabet & Boughazi, 2011). Some of the earliest effective speech synthesis models were made using a parametric approach which requires the careful setting of speech parameters in order to synthesize the necessary sounds in a language (Black et al., 2007). The model is fed text to inform the system which sounds are to be produced and in which order. Concatenative speech synthesis is done by cutting speech segments from existing recordings and storing them with text labels (Hunt & Black, 1996). Speech is then synthesized by splicing the corresponding audio snippets in the order designated by a text input. Parametric and concatenative methods can produce high quality results but require extensive domain expertise and manual preparations. The leading method of synthesizing speech is to do so by training a deep neural network (DNN) (Ning et al., 2019). The DNN approach is able to produce

industry-leading synthesized speech and, in the case of frameworks such as Tacotron (Wang et al., 2017), can do so without requiring the annotation of data. These advantages come at a cost, however, as the DNN approach requires much more data than other methods.

Tacotron has become an extremely popular deep neural network framework for TTS applications because of its ability to use graphemes as training input and because it exhibits minimal pronunciation errors in the output speech (Perquin et al., 2020). Using phoneme inputs has proven to be a viable option which may improve results but also requires much more preprocessing of the text and linguistic expertise to execute. Many studies have shown that Tacotron is able to model languages with various characteristics (Kwon et al., 2020; Nthite & Tsoeu, 2020; Yang et al., 2019; Zhang et al., 2019). The fact that Tacotron can use raw audio and text transcriptions as training data make it a much more accessible option for an under-resourced language like Malagasy, for which aligned transcriptions are largely unavailable. Additionally, unless the user intends to alter Tacotron's architecture, training Tacotron models does not require expertise in linguistics and requires only a limited understanding of machine learning. This is advantageous as it expands the number of people who could train a model in their native language. For these reasons, I chose to use the Tacotron framework to develop TTS models in the present study. Specifically, I use an open-access version called Tacotron 2, which is developed by the Nvidia corporation (*Tacotron 2 (without Wavenet)*, 2018/2022).

Figure 1 shows that Tacotron is an "end-to-end" system because it can produce an audible speech waveform (seen at the end of the pipeline) from a pure text input (fed in at the start of the pipeline). The various layers of the neural network are used to learn associations between the text and its corresponding speech.

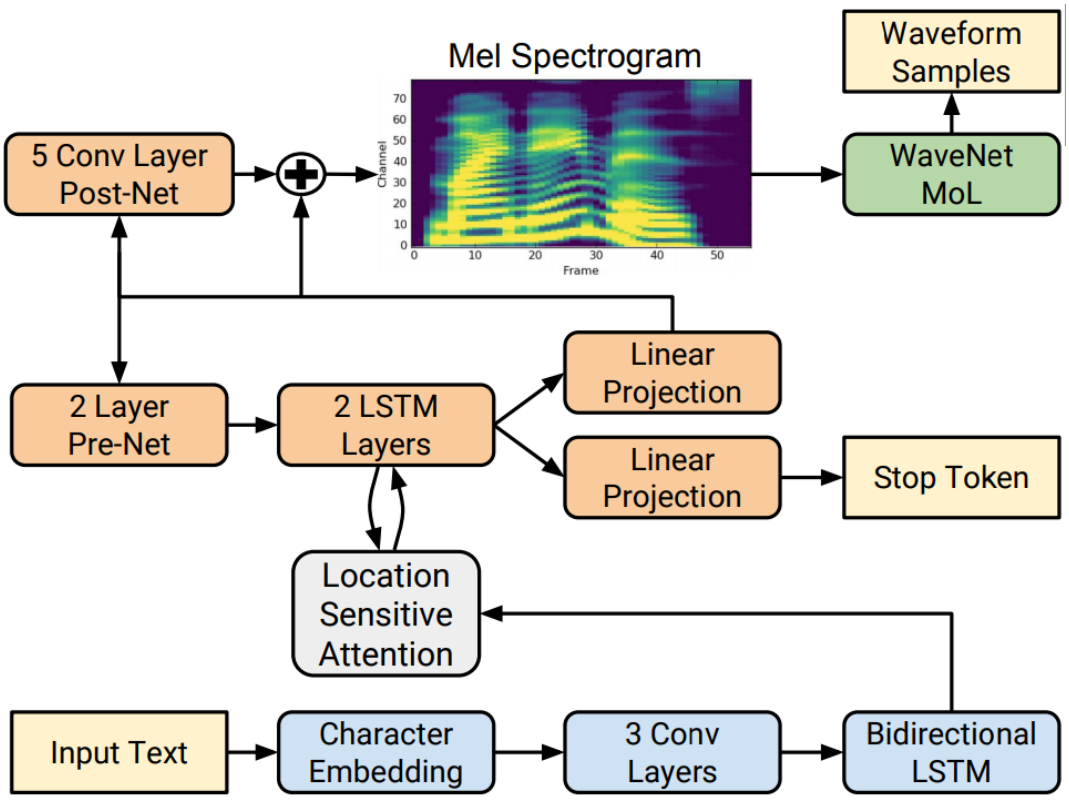

Figure 1: The architecture of Tacotron 2 (Shen et al., 2018)

3.1 Development Process

As mentioned in section 1.1, one of the objectives of the present study is to explore the effects of methods which might facilitate the development of speech synthesis models for under-resourced languages. I strive to systematically test these methods by training models using the data described in chapter two. One of my goals for this part of the project is to document these methods so that they are replicable by any who have the necessary programming skills, regardless of whether they have domain knowledge in linguistics. With each new method used and model trained I note the general changes to the output of the model. I use this stepwise

approach so that those who intend to build models for other languages may be informed as to what works and what does not. I also give my opinion regarding which steps have the largest impact on the development of the model so that those who have limited resources to devote to their models may know where to focus their efforts.

There are limitations to this approach, as I can only give my own general observations of differences between models. As a non-native speaker of Malagasy, my observations are impressionistic and are not generalizable to what may be experienced by native speakers. When I evaluate the intelligibility of the model's productions, I justify my claims to the best of my ability by discussing differences between the expected speech sounds and those produced by the model. Likewise, I use pitch contours and spectrograms to justify my evaluation of voice quality and human-likeness. Unfortunately, quantitative evaluation by native speakers of each of the models is impossible due to the time constraints of the present study. Evaluation by native speakers will be carried out after models are refined in a future study. This will be discussed in more detail in section 4.1. Despite these limitations, I argue that the following models provide general indications as to which methods are most effective for developing TTS models for under-resourced languages.

3.2 Models

3.2.1 No Alterations

I first trained a baseline model with no alterations. I did this in order to determine whether the unmodified framework and only the single-speaker data would be enough to train an

effective model. Given my aim of taking practicality and domain expertise into account, an unaltered dataset would be ideal. If the engineer was able to simply organize the small dataset and train the model using a single command, the process would be trivial to replicate for any number of languages. Additionally, using the default training parameters (the English model training parameters) means less work and expertise required of the engineer. This is not to say that the training parameters found best by the developers of Tacotron 2 (2018/2022) for training on a large English dataset are best for modeling every language or datasets of every size. However, it is useful for engineers to know whether the adjustment of training parameters is strictly necessary in order to achieve a practical TTS model using a small speech dataset.

Unfortunately, but not unexpectedly, this approach does not work. Training a deep neural network on a half hour of single-speaker data does not present the network with enough exposure to text-speech pairs in order to learn associations between them. The output of this baseline model is speechlike but completely unintelligible. There appears to be little to no correlation between the text provided as input to the model and the synthesized sounds. The spectrogram shown in Figure 2 indicates a relative lack of information in the mid-high frequency regions of some sections. Furthermore, the pitch contour in the same figure is much more erratic than what is typically observed in speech. The failure of this model illustrates the main obstacle in developing a TTS for an under-resourced language using deep neural networks: a large amount of data is required in order to be successful.

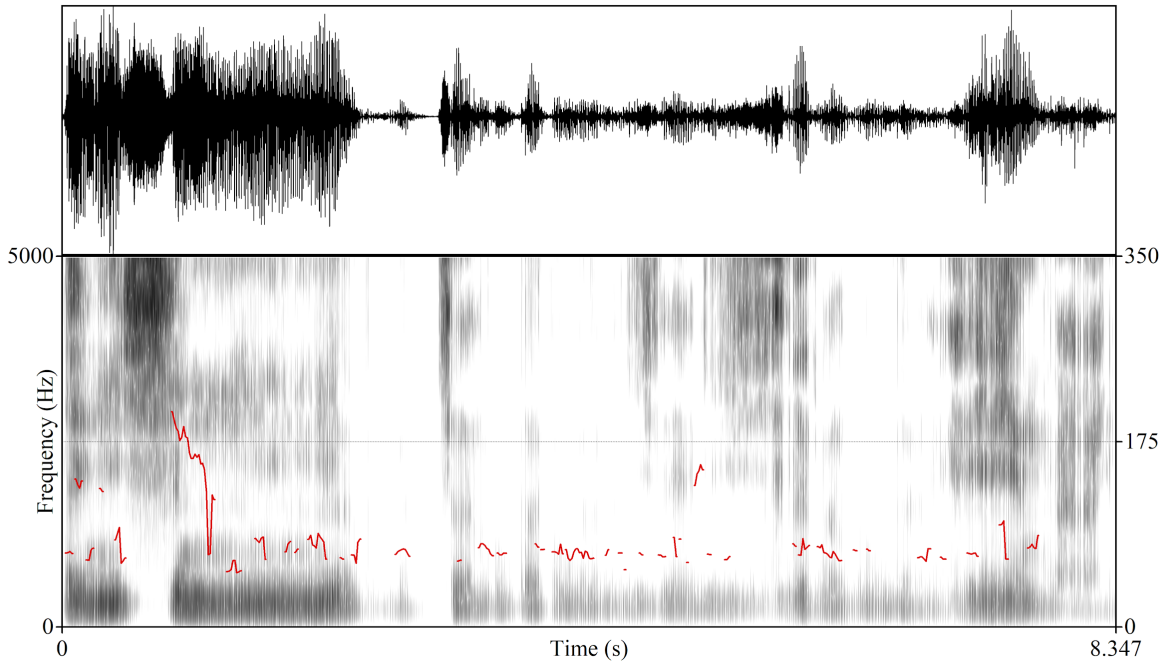

Figure 2: A waveform and spectrogram of this model's synthesis of the phrase, "Ny Fanahy sy ny ampakarina" ("The Spirit and the bride" in English). The pitch contour (shown in red over the spectrogram) was calculated using a range from 50 to 350 Hz. Word boundaries are not labeled in this figure because the speech is unintelligible. The large gaps in information at certain frequencies, the ill-defined speech segments, and the erratic pitch which are evident in this figure all lend to the unintelligible nature of this model's output.

3.2.2 Cross-Lingual Transfer Learning

One common approach taken by machine learning engineers is to use transfer learning to overcome data deficits (Zhuang et al., 2021). By training a source model on a large dataset in a different language, associations between graphemes and sounds which are relevant in both that

language and the target language may be learned. Even if the production of certain sounds in the target language are different from the language which is used for preliminary training, this training gives a better foundation for the model than the random associations used as a starting point if there is no transfer learning.

I used the published English Tacotron 2 model as the source model for training the Malagasy target model. It was trained using the LJ Speech dataset, which is composed of approximately 24 hours of short clips from passages of reading by a single speaker. I will henceforth refer to the LJ Speech Dataset and the model on which it is trained as the English dataset and the English model or the source dataset and the source model.

The English speaker has very different speech characteristics from the single-speaker in the Malagasy dataset. The average fundamental frequency of the Malagasy speaker is over 100 hertz lower than the English speaker, for example. Not only is the source and target data different at the speaker level, but differences in the sound inventories of English and Malagasy mean that there is also a divide at the language level. Malagasy contains sounds which are not produced in English, like prenasalized stops (e.g. /ᵐb/ and /ⁿt/) and prenasalized affricates (e.g. /ⁿt͡ʃ/ and /ⁿ͡ɖʐ/). Additionally, North American English contains sounds which are not produced in Malagasy, like the alveolar approximant /ɹ/.

Despite differences between the source and target data, transfer learning greatly improved the model's output. Compared to the baseline model, the transfer learning model has improved intelligibility. Evidence for the improved intelligibility may be found in the spectrogram presented in Figure 3, where the characteristics of each speech segment are generally consistent with those expected given the text input. Formants are distinguishable in the synthesized vowels,

each speech segment has a proportionate duration within the range one might expect from human speech, and stop closures are relatively free from noise. However, the still limited exposure to Malagasy speech causes the model to make errors. For example, the production shown in Figure 3 has an unnatural voice quality which is similar to creaky voice (Keating et al., 2015). Evidence of this can be seen in the pitch contour of Figure 3. Although the contour is much smoother than Figure 2, the unexpected breaks and high variation in pitch still lend to speech which sounds less natural. Additionally, the output of this model sounds less human-like because of high levels of noise.

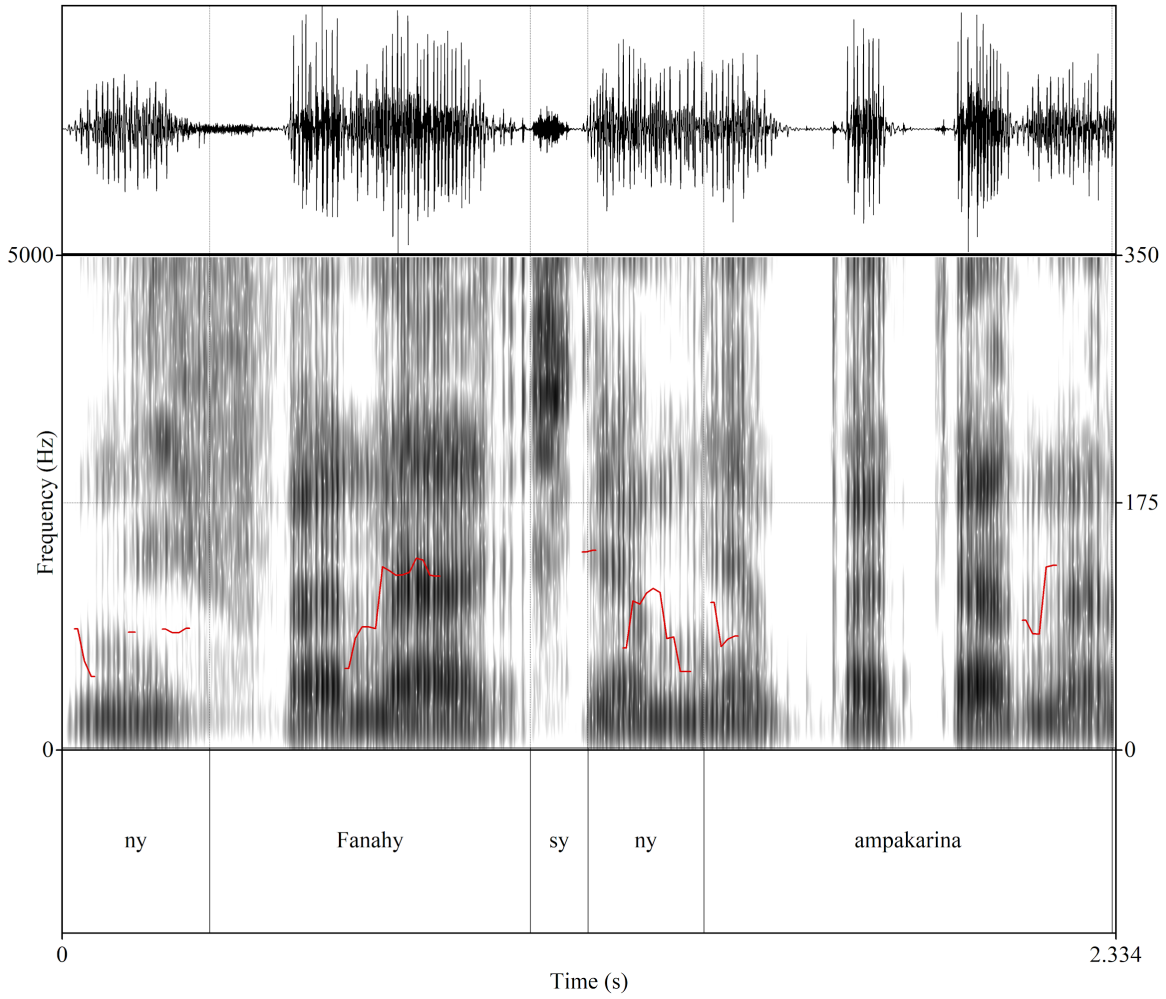

Figure 3: A waveform and spectrogram of this model's synthesis of the phrase, "Ny Fanahy sy ny ampakarina" ("The Spirit and the bride" in English). The pitch contour (shown in red over the spectrogram) was calculated using a range from 50 to 350 Hz. Word boundaries are indicated with a text-grid. There are notable improvements visible in the spectrogram, like distinguishable formants and other identifiable segment characteristics. The pitch contour is less erratic than that shown in Figure 2, but still shows high variability and unnatural transitions from segment to segment.

3.2.3 Resampling

Resampling the Malagasy dataset to be consistent with the English dataset was also found to be an important step for improving the model's output. After resampling the Malagasy data to the rate of the English data (22050 Hz), model outputs were improved again. The most notable improvements include a change in overall voice quality, which became more clear and less noisy. The reduction in noise is evident when comparing features which are typically silent or nearly silent. For example, the stop closures in Figure 4 are slightly less noisy than those in Figure 3. The prosody of the speech was also affected, and more closely matches the cadence of human speech after resampling. The more human-like speech rate in this model's output is the most notable cause of this. The unnaturally slow speech rate produced by the model described in 3.2.2 is a result of training on data which features an inconsistent number of samples per amount of time. Training on data which has a consistent sampling rate allows the model to learn the appropriate durations of speech sounds. Overall, resampling made the synthesized speech more intelligible and human-like.

Although the output's prosody is improved by training on resampled data, the model still makes occasional errors. The most notable error is the addition of a strange speech sound which does not correspond to the text input. An example of this can be seen in Figure 4, in which an added segment can be seen even after speech corresponding to all of the text has been synthesized. These strange additions are a result of inadequate exposure to Malagasy text-speech pairs. Additionally, the synthesized speech still exhibits a somewhat raspy tone with uneven, wavering pitch. This is apparent in the pitch contour shown in Figure 4. These errors are also assumed to be a result of the model's limited exposure to Malagasy text-speech pairs. Sections 3.2.4 and 3.2.5 will describe two attempts to maximize the model's exposure to Malagasy data in order to improve robustness.

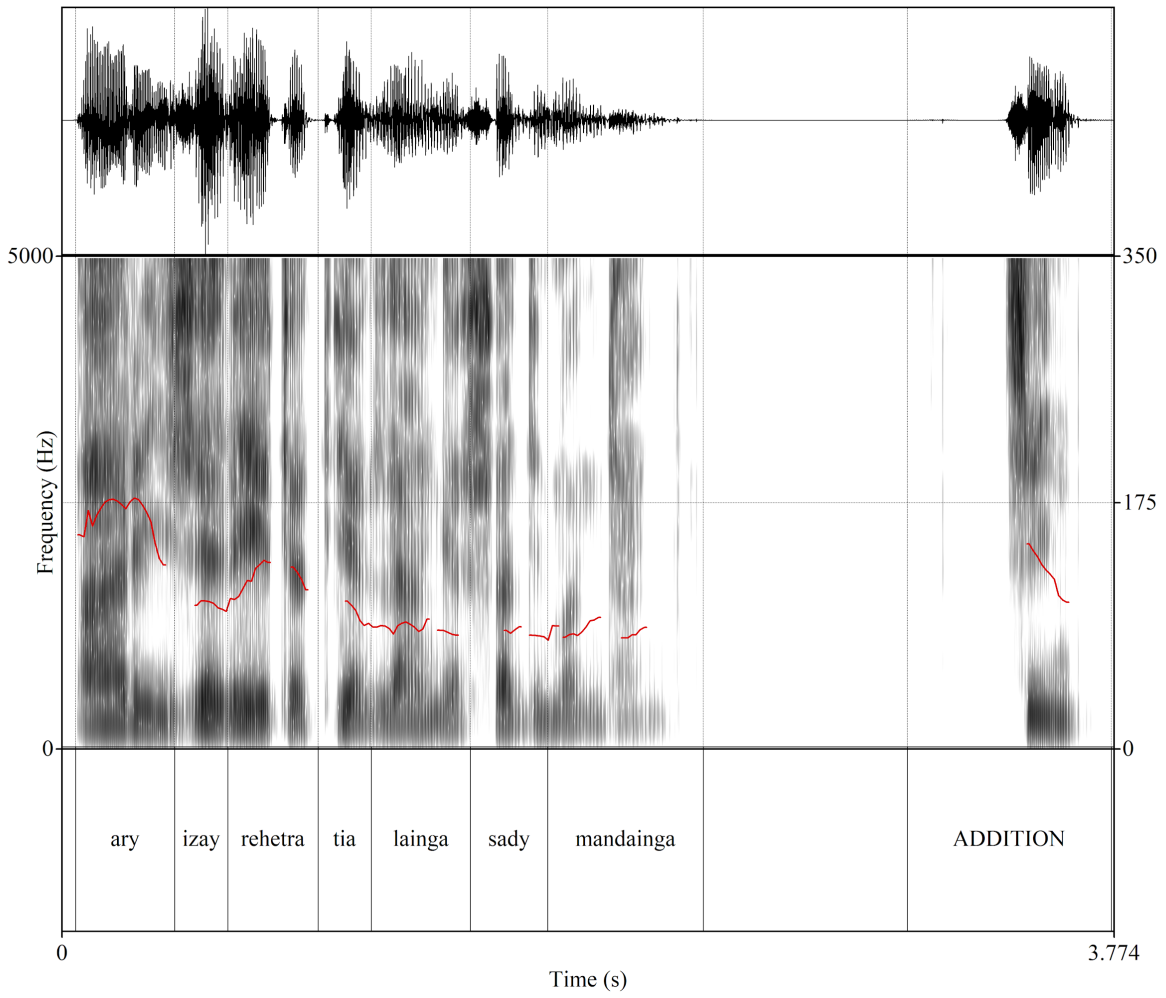

Figure 4: A waveform and spectrogram of this model's synthesis of the phrase, "ary izay rehetra izay tia lainga sady mandainga" ("whosoever loveth and maketh a lie" in English). The pitch contour (shown in red over the spectrogram) was calculated using a range from 50 to 350 Hz. Word boundaries are indicated with a text-grid. This model synthesizes speech at a more natural speech rate but the speech still includes the occasional addition of strange sounds (see the ADDITION label for an example).

3.2.4 Augmentation

The augmentation of data allows a machine learning engineer to increase a model's exposure to the desired text-speech pairs without the need for new recordings. This may be done by the addition to recordings of levels of noise which do not obscure the speech but place it in a slightly different environment. Training the model on the same speech sounds with slightly varying levels of noise makes the model's associations between text and speech more robust.

I added 40 dB white noise to the training recordings using Praat (Boersma & Weenink, 2022) and the Praat Vocal Toolkit (Corretge, 2012) to make an augmented copy of each file. It is my impression that the model trained on the augmented files produces higher quality synthesized speech compared to versions without augmented training data. Limitations in the performance of neural network models are most often due to the size of the training dataset. Small datasets do not provide models with enough exposure to the data in order to learn the associations between text-speech pairs. Therefore, it is logical that the data augmentation strategy described above—which effectively doubles the amount of training data—improves the model's outputs. Augmentation techniques are a productive and simple way of improving the effectiveness of small speech datasets for training neural networks.

Improvements to the voice quality are evident in the smoothness of the pitch contour shown in Figure 4. Additionally, the speech segments in Figure 4 are better defined than the previous models' productions. Stops feature little noise during closures, fricatives—like the utterance-final fricative in Figure 4—exhibit noise concentrated in the appropriate frequencies, and formants are distinguishable in the vowels.

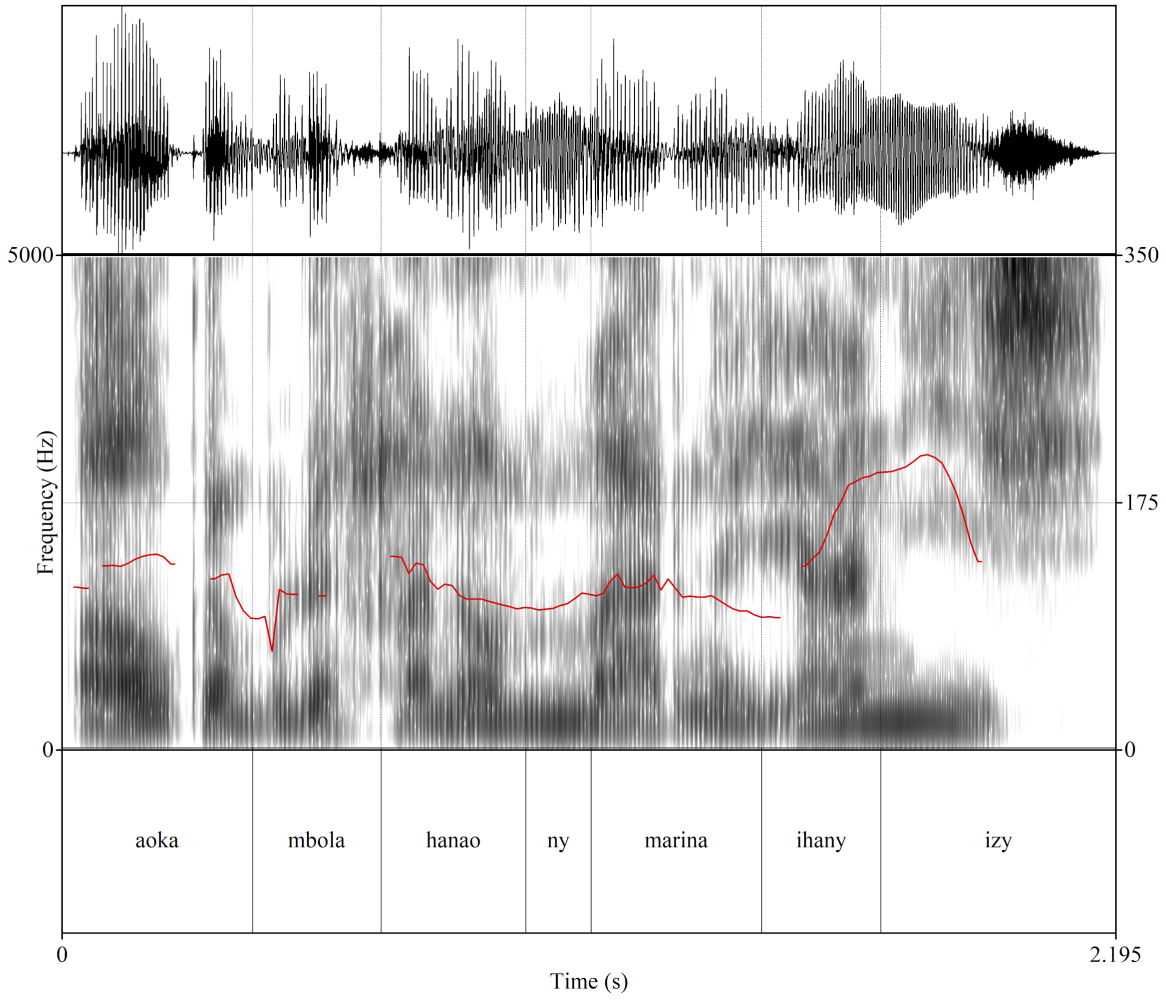

Figure 5: A waveform and spectrogram of this model's synthesis of the phrase, "Aoka mbola hanoa ny marina ihany izy" ("let him be righteous still" in English). The pitch contour (shown in red over the spectrogram) was calculated using a range from 50 to 350 Hz. Word boundaries are indicated with a text-grid. The pitch contour in this model's output features more smooth, human-like transitions than previous models. Speech segments are more easily recognizable and more clearly correspond to the text inputs than other models.

3.2.5 Combination of Single-Speaker and Crowd-Sourced Data

Speech synthesis which is somewhat practical can be achieved using a combination of cross-lingual transfer learning; a small, single-speaker speech dataset in the target language; and some simple data manipulation techniques. However, as mentioned in section 2.2, there can be advantages to using crowd-sourcing methods to collect a training dataset. In some cases, using a small amount of data from a large number of speakers may be the only currently available resource for use in training a TTS model. So, despite having achieved promising results using only the single-speaker data, I thought it valuable to do some preliminary testing using the crowd-sourced data.

This model is the same as the model described in section 3.2.4—it uses the same English source model introduced in section 3.2.2, all recordings are resampled to be consistent with the source dataset, and augmented versions are made of each file—except the target dataset consists of both the single-speaker data (~0.5 hours) and the crowd-sourced data (~0.8 hours). I decided to include both single-speaker and crowd-sourced data in the training dataset for this test because the latter alone would not give enough exposure to the model of a single, consistent speaking style, likely leading to an unintelligible output.

Outputs from this model suggest that adding crowd-sourced data to a small, single-speaker training dataset decreases the effectiveness of the model significantly. This instance of the model does not produce intelligible speech and the associations between graphemes and speech sounds appear to be extremely confused. In earlier models—like the one described in section 2.2—the appropriate sounds are made for the given inputs even if the voice quality is not very human-like. The addition of multispeaker data to this model's training dataset

appears to have broken down those text-speech associations. The sounds which are produced are fairly noise-free and are speechlike, but do not correspond to the text input. Additionally, outputs include inappropriate pauses and repetitions of small speech segments at unexpected points in the utterance, as can be seen in Figure 6.

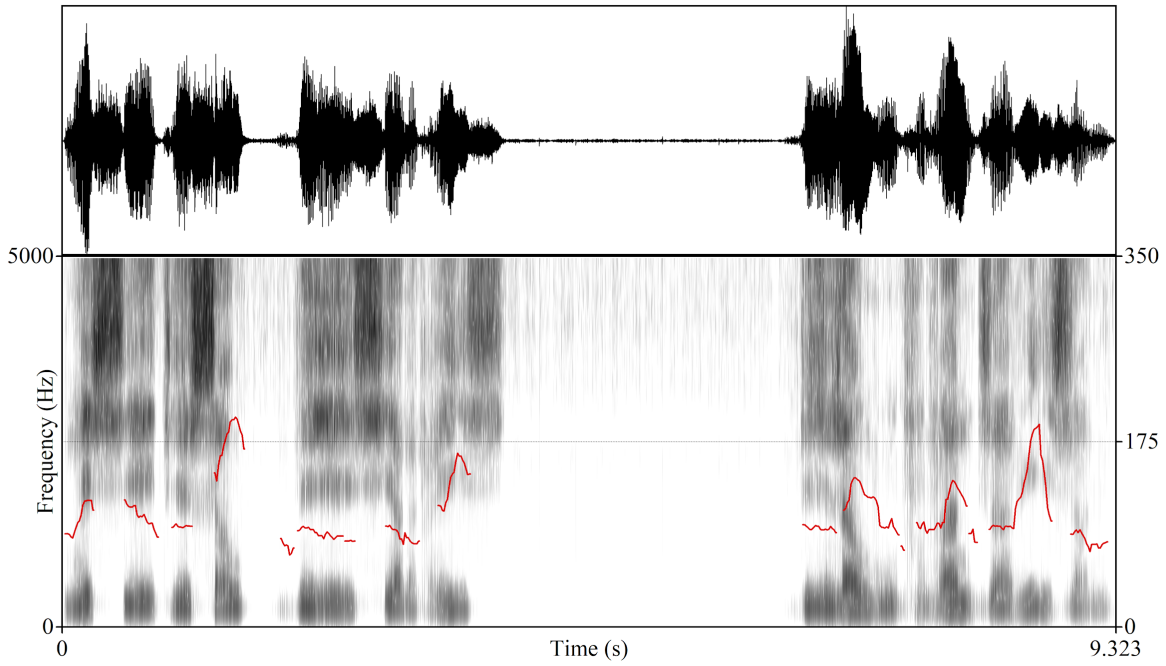

Figure 6: A waveform and spectrogram of this model's synthesis of the phrase, "Ny Fanahy sy ny ampakarina" ("The Spirit and the bride" in English). The pitch contour (shown in red over the spectrogram) was calculated using a range from 50 to 350 Hz. Word boundaries are not labeled in this figure because the speech is unintelligible. The pitch contours are more human-like than the initial model's output and the segments are well defined and relatively free of noise. However, the synthesized sounds do not correspond to the text input, indicating that the model's associations of Malagasy text and speech have been confused by the addition of multi-speaker training data.

As previously mentioned, introducing multi-speaker data to a TTS framework is not ideal because of the high amount of variance exhibited by said data. The crowd-sourced data used in this study may even be an extreme example of this variance as it was collected remotely from a wide range of recording environments. Especially when it comes to the data-hungry nature of neural network architectures and the small datasets which must be used when working with under-resourced languages, variance can be a considerable obstacle. This does not necessarily mean that crowd-sourced data can never benefit these sorts of models, as many factors may have contributed to the decrease in quality. One probable cause is the quantity of crowd-sourced data relative to the quantity of single-speaker data. It could be that training a model with a higher ratio of single-speaker to crowd-sourced data would allow the data from multiple speakers to contribute to the model's associations between text and speech without overriding and confusing the characteristics of the single speaker. The inclusion of multi-speaker data in the source dataset may also present a more effective way to give a model more exposure to phonetic productions which may be specific to the target language while allowing the target dataset to consist of only consistent productions by a single speaker. The implications of using a mixed source dataset and further discussion on the use of crowd-sourced data will be included in section 4.1.

CHAPTER IV. Discussion and Conclusions

Chapter four begins with a summary and discussion of the speech synthesis results from chapter three. Then, I address the implications of the present study's findings for those who are interested in TTS for under-resourced languages. In closing, weaknesses in the present study and ideas for future research are described.

The first research objective of this project was to use the Tacotron TTS framework to create an effective Malagasy speech synthesis model. Although native speaker perceptions of the intelligibility and naturalness of the synthesized speech would be a better measure of success for this objective, my own intuition is that the better-performing models synthesize Malagasy speech well enough to be useful. There are relatively few errors in the synthesized test phrases, the voice quality is somewhat natural, and the speech is intelligible in most instances.

Each of the models described in section 3.2 were trained with controlled differences between each step. Each successive model features an additional method by which I attempt to improve the model's output. With the exception of the model which was trained on a combination of single-speaker and crowd-sourced data, all of the methods described above improved the model in some way. The most significant improvement is found when cross-lingual transfer learning is implemented. This improvement is a result of the target's use of the associations learned from many hours of English data as a starting point instead of random weights. There is value in noting that although there are many similarities between English and Malagasy productions which correspond to the same graphemes, there are also marked differences. Metrics which have historically been used to determine which source language should be used for cross-lingual transfer learning are mainly arbitrary, the most common measure

being language family (Do et al., 2021). English and Malagasy are unrelated according to this measure, meaning my findings support those of Do et al. who do not find a significant relationship between language family and the effectiveness of models trained using cross-lingual transfer learning. The steps involving resampling and augmentation either enhanced the effectiveness of the cross-lingual transfer or served to make the most of the small target dataset by increasing consistency or exposure. Efforts to combine crowd-sourced and single-speaker data in a single target set did not improve the model's output at their current ratio.

4.0 Implications

The results of this project have implications for those who want to develop TTS models for under-resourced languages. First, the collection of crowd-sourced speech data should not be prioritized until a considerable amount of single-speaker data is available for training. Multi-speaker data may benefit the model, but further testing is needed before it can be determined whether the multi-speaker data must be limited to the source dataset. It may also be that multi-speaker data can be combined with single-speaker target data in a ratio that will not override the characteristics of the predominant speaker while still being beneficial. Multi-speaker data may provide additional exposure to text-speech associations in much the same way the source language does, so a large multi-speaker dataset may be an effective alternative to using a source dataset/model from another language. In any case, single-speaker data is essential to the TTS model's success and should be prioritized.

A second implication is that engineers need not limit themselves to using source datasets from the same language family as the target language. The present study is an example of

success using unrelated languages and Do et al. (2021) do not find a significant difference in their systematic review.

4.1 Future Work

As discussed in section 2.1, taking the crowd-sourcing approach to data collection presents several challenges relating primarily to a lack of control over the recording quality. In order to improve the effectiveness of the crowd-sourcing approach one must implement methods of data verification and participant instruction which are ultimately aimed at improving data quality. For example, machine learning and other approaches to outlier detection can perform with extremely high accuracy (Larson et al., 2019; Poorjam et al., 2019; Zhao & Hryniewicki, 2018). In an especially relevant study, Zequeira Jiménez et al. (2018) investigate the efficacy of outlier detection and control questions in the specific context of a crowd-sourced speech dataset. Future work may include a replication of the methods described by Zequeira Jiménez et al. on the crowd-sourced Malagasy dataset and investigation to determine whether said methods improve the efficacy of speech synthesis models which are trained on the cleaned data.

A deeper investigation regarding the effects of two data ratios would also answer many of the questions brought on by the present study. First, the ratio of crowd-sourced target data to single-speaker target data is of interest because the results of the present study suggest that a ratio of 0.78:0.50 hours is ineffective. The second ratio is the ratio of source data to single-speaker target data. In the present study I found that using the LJSpeech dataset as the source and a small single-speaker Malagasy dataset as the target (a ratio of 24:0.5 hours) was effective. However, the development of speech synthesis models for under-resourced languages

may be more accessible if it is possible to use a smaller source dataset in cross-lingual transfer learning while preserving the quality of the synthesized speech. A determination of the optimal ratios in both cases would require more data and the training and comparison of models with a range of data ratios. Additionally, future research should be conducted using a single-speaker dataset which is recorded specifically for use in the development of speech technologies. This would allow for greater control over the data and make findings easier to disseminate.

An additional area of interest is the effectiveness of including crowd-sourced data from the target language in the source dataset. I predict that providing early exposure of the source model to the target language would improve the robustness of the target model's output without confusing it by including multiple speaker styles in the target dataset. Again, a range of data ratios (source language:crowd-sourced target language data) should be tested in order to determine the composition of the optimal source dataset. The inclusion of crowd-sourced data in the source dataset should have been tested first in the present study because I predict it would have had a less pronounced impact on the target model's outputs compared to inclusion in the target data.

The use of different source languages for cross-lingual transfer learning has already been examined in some respects (Do et al., 2021) and the results of the present study suggest that languages from different language families can be used to train effective speech synthesis models. However, I believe it would still be valuable to train and compare models with different source languages. As language family has not been shown to have a significant effect, other measures such as phoneme inventory likeness between source and target languages may be a more appropriate way to predict optimal source languages. Further investigation is also needed to

determine whether similarity between source and target data at the speaker level has a significant effect.

If the effectiveness of using crowd-sourced speech data is truly limited by the variation of speakers, it may be that methods of speaker normalization would negate some of that variation while still preserving valuable information for the model. In the context of the present study, speaker normalization refers to an effort to manually reduce the amount of variation between the speech produced by different speakers in order to create more consistent training data. For example, efforts to develop automatic speech recognition models for children's speech have been improved by applying vocal tract length normalization (VTLN) to training data (Qian et al., 2016). VTLN has also been used to make adaptive parametric speech synthesis (Saheer et al., 2014). As a simplified preliminary effort, I attempted to use Praat's (Boersma & Weenink, 2022) PSOLA algorithm to bring the fundamental frequency of the target speech up to the pitch of the source speaker. This resulted in an unnatural voice tone in the synthesized speech and did not appear to improve the output. It quickly became apparent that it was not essential to normalize the Malagasy target speech, as the target speech style generally overrode the characteristics of the source speech style. However, it is yet to be determined whether the normalization of the crowd-sourced speech data would make its inclusion in the target dataset more effective. A survey of the available methods of speaker normalization for TTS data would be valuable, as would testing to determine which is most effective.

Presently, all evaluations of the Malagasy speech synthesis model's performance consist of qualitative evaluation by myself (a non-native Malagasy speaker). Future work could include testing like that done by Nthite and Tsoeu (2020) where native speakers are asked to transcribe synthesized speech and rate its naturalness. Such quantitative measures of intelligibility,

accuracy, and naturalness by native speakers would give more precise information about which changes to the data and model were most beneficial.

4.2 Conclusion

The data collection portion of the present study focused on the use of an application I created to crowd-source a small Malagasy speech dataset. The crowd-sourced dataset was used in conjunction with and in comparison to a small single-speaker dataset which was gathered from audiobook recordings online. Both datasets were used and various methods were tested in order to determine their effects on the final product. The inclusion of both crowd-sourced and single-speaker data in a target dataset was found to confuse the model at the ratio used in the present study, but further investigation is required in order to determine whether crowd-sourced data may be useful in different ratios or as part of the source dataset. Cross-lingual transfer learning was determined to be extremely useful in improving model outputs and, in support of the findings of Do et al. (2021), the fact that English is not related to Malagasy was not found to preclude it from being used as an effective source language. Resampling and data augmentation were also found to improve model outputs. Further investigation and evaluation by native speakers will determine whether the goal of training a practical Malagasy TTS model was achieved.

Machine learning approaches to speech synthesis are currently unmatched in terms of output quality, but require large amounts of training data in order to be effective. This presents difficulty when developing models for under-resourced languages, for which data is scarce. The findings of the present study suggest that even very small speech datasets can be leveraged in a number of ways to train effective models. The methods by which this can be done may prove

valuable to language communities which have gone without the benefits of speech technologies due to data scarcity. Especially in the light of increasing rates of language endangerment (Bromham et al., 2022), effort should be made to create speech corpora which can be used for language documentation and description as well as the development of speech technologies. In cases where small speech datasets are already available in a language of interest, the methods discussed in the present study may be applied in order to develop a TTS model in that language. The increased availability of speech technologies for under-resourced languages may offer significant benefits to speakers of those languages.

REFERENCES

Adelaar, A. (2018). Seventeenth century texts as a key to Malagasy linguistic and ethnic history. *Proc. 14th International Conference on Austronesian Linguistics*, 17–20.

Anumanchipalli, G. K., & Black, A. W. (2010). Adaptation techniques for speech synthesis in under-resoured languages. *Proc. 2010 Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU)*, 51–55. https://www.cs.cmu.edu/afs/cs.cmu.edu/Web/People/gopalakr/publications/sltu2010_anumanchipalli.pdf

Aziz, J. (2020). Intonational Phonology of Malagasy: Pitch Accents Demarcate Syntactic Constituents. *Proc. Speech Prosody 2020*, 201–204. https://doi.org/10.21437/SpeechProsody.2020-41

Aziz, J., & Paul, I. (2019). The Intonation of Malagasy: A preliminary look. *Proc. the 19th International Congress of the Phonetic Sciences*, 3792–3796.

Besacier, L., Barnard, E., Karpov, A., & Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. *Speech Communication*, *56*, 85–100. https://doi.org/10.1016/j.specom.2013.07.008

Blachon, D., Gauthier, E., Besacier, L., Kouarata, G.-N., Adda-Decker, M., & Rialland, A. (2016). Parallel Speech Collection for Under-resourced Language Studies Using the Lig-Aikuma Mobile Device App. *Procedia Computer Science*, *81*, 61–66. https://doi.org/10.1016/j.procs.2016.04.030

Black, A. W., Zen, H., & Tokuda, K. (2007). Statistical Parametric Speech Synthesis. *Proc. 2007 IEEE International Conference on Acoustics, Speech and Signal Processing - ICASSP '07*, *4*, IV-1229-IV–1232. https://doi.org/10.1109/ICASSP.2007.367298

Boersma, P., & Weenink, D. (2022). *Praat: Doing phonetics by computer* (6.2.12) [Computer software]. http://www.praat.org/

Bromham, L., Dinnage, R., Skirgård, H., Ritchie, A., Cardillo, M., Meakins, F., Greenhill, S., & Hua, X. (2022). Global predictors of language endangerment and the future of linguistic diversity. *Nature Ecology & Evolution*, *6*(2), 163–173. https://doi.org/10.1038/s41559-021-01604-y

Brucato, N., Kusuma, P., Cox, M. P., Pierron, D., Purnomo, G. A., Adelaar, A., Kivisild, T., Letellier, T., Sudoyo, H., & Ricaut, F.-X. (2016). Malagasy Genetic Ancestry Comes from an Historical Malay Trading Post in Southeast Borneo. *Molecular Biology and Evolution*, *33*(9), 2396–2400. https://doi.org/10.1093/molbev/msw117

Buddhika, D., Liyadipita, R., Nadeeshan, S., Witharana, H., Jayasena, S., & Thayasivam, U. (2018). Voicer: A Crowd Sourcing Tool for Speech Data Collection. *Proc. 18th International Conference on Advances in ICT for Emerging Regions (ICTer)*, 174–181. https://doi.org/10.1109/ICTER.2018.8615521

Chen, S. S., Eide, E., Gales, M. J. F., Gopinath, R. A., Kanvesky, D., & Olsen, P. (2002). Automatic transcription of Broadcast News. *Speech Communication*, *37*(1), 69–87. https://doi.org/10.1016/S0167-6393(01)00060-7

Corretge, R. (2012). *Praat Vocal Toolkit*. http://www.praatvocaltoolkit.com

de Vries, N. J., Davel, M. H., Badenhorst, J., Basson, W. D., de Wet, F., Barnard, E., & de Waal, A. (2014). A smartphone-based ASR data collection tool for under-resourced languages. *Speech Communication*, *56*, 119–131. https://doi.org/10.1016/j.specom.2013.07.001

Do, P., Coler, M., Dijkstra, J., & Klabbers, E. (2021). A Systematic Review and Analysis of Multilingual Data Strategies in Text-to-Speech for Low-Resource Languages.

Interspeech 2021, 16–20. https://doi.org/10.21437/Interspeech.2021-1565

Dumitrescu, S. D., Boroș, T., & Ion, R. (2014). Crowd-sourced, automatic speech-corpora collection–Building the Romanian Anonymous Speech Corpus. *CCURL 2014: Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era*, 90–94.

Farner, S., Veaux, C., Beller, G., Rodet, X., & Ach, L. (2008). *Voice transformation and speech synthesis for video games*. 3.

Free Speech... Recognition (Linux, Windows and Mac)—Voxforge.org. (2006, 2022). VoxForge. http://www.voxforge.org/home

Freitas, D. (2010). Accessibility and Design for All Solutions Through Speech Technology. In F. Chen & K. Jokinen (Eds.), *Speech Technology: Theory and Applications* (pp. 271–299). Springer US. https://doi.org/10.1007/978-0-387-73819-2_14

Hammarström, H., Forkel, R., Haspelmath, M., & Bank, S. (2021). *Glottolog 4.5* [Data set]. Zenodo. https://doi.org/10.5281/zenodo.5772642

Higginbotham, D. J. (2010). Humanizing Vox Artificialis: The Role of Speech Synthesis in Augmentative and Alternative Communication. In J. Mullennix & S. Stern (Eds.), *Computer Synthesized Speech Technologies: Tools for Aiding Impairment* (pp. 50–70). IGI Global. https://doi.org/10.4018/978-1-61520-725-1.ch004

Hunt, A. J., & Black, A. W. (1996). Unit selection in a concatenative speech synthesis system using a large speech database. *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, *1*, 373–376 vol. 1. https://doi.org/10.1109/ICASSP.1996.541110

Keating, P., Garellek, M., & Kreiman, J. (2015). Acoustic properties of different kinds of creaky

voice. *ICPhS*, *2015*(1), 2–7.

Kwon, M., Jeong, Y., & Choi, H. (2020). Implementation of Python-Based Korean Speech Generation Service with Tacotron. *2020 IEEE International Conference on Big Data and Smart Computing (BigComp)*, 551–552. https://doi.org/10.1109/BigComp48618.2020.000-5

Larson, S., Mahendran, A., Lee, A., Kummerfeld, J. K., Hill, P., Laurenzano, M. A., Hauswald, J., Tang, L., & Mars, J. (2019). Outlier Detection for Improved Data Quality and Diversity in Dialog Systems. *ArXiv:1904.03122 [Cs]*. http://arxiv.org/abs/1904.03122

Li, S., Xu, L. D., & Zhao, S. (2015). The internet of things: A survey. *Information Systems Frontiers*, *17*(2), 243–259. https://doi.org/10.1007/s10796-014-9492-7

Liao, H., McDermott, E., & Senior, A. (2013). Large scale deep neural network acoustic modeling with semi-supervised training data for YouTube video transcription. *Proc. 2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, 368–373. https://doi.org/10.1109/ASRU.2013.6707758

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, *54*(6), 115:1-115:35. https://doi.org/10.1145/3457607

Mills, T., Bunnell, H. T., & Patel, R. (2014). Towards Personalized Speech Synthesis for Augmentative and Alternative Communication. *Augmentative and Alternative Communication*, *30*(3), 226–236. https://doi.org/10.3109/07434618.2014.924026

Ning, Y., He, S., Wu, Z., Xing, C., & Zhang, L.-J. (2019). A Review of Deep Learning Based Speech Synthesis. *Applied Sciences*, *9*(19), 4050. https://doi.org/10.3390/app9194050

Nthite, T., & Tsoeu, M. (2020). End-to-end text-to-speech synthesis for under resourced South

African languages. *2020 International SAUPEC/RobMech/PRASA Conference*, 1–6. https://doi.org/10.1109/SAUPEC/RobMech/PRASA48453.2020.9041030

Ny Baiboly malagasy. (n.d.). Henoy Ny Baiboly. Retrieved April 27, 2022, from https://nybaiboly.net/

Panayotov, V., Chen, G., Povey, D., & Khudanpur, S. (2015). Librispeech: An ASR corpus based on public domain audio books. *Proc. 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5206–5210. https://doi.org/10.1109/ICASSP.2015.7178964

Parent, G., & Eskenazi, M. (2011). Speaking to the Crowd: Looking at Past Achievements in Using Crowdsourcing for Speech and Predicting Future Challenges. *Proc. Interspeech 2011*, 2027–3040. https://www.isca-speech.org/archive_v0/archive_papers/interspeech_2011/i11_3037.pdf

Perquin, A., Cooper, E., & Yamagishi, J. (2020). Grapheme or phoneme? An Analysis of Tacotron's Embedded Representations. *ArXiv:2010.10694 [Cs]*. http://arxiv.org/abs/2010.10694

Polyakov, E. V., Mazhanov, M. S., Rolich, A. Y., Voskov, L. S., Kachalova, M. V., & Polyakov, S. V. (2018). Investigation and development of the intelligent voice assistant for the Internet of Things using machine learning. *Proc. 2018 Moscow Workshop on Electronic and Networking Technologies (MWENT)*, 1–5. https://doi.org/10.1109/MWENT.2018.8337236

Poorjam, A. H., Little, M. A., Jensen, J. R., & Christensen, M. G. (2019). Quality Control in Remote Speech Data Collection. *IEEE Journal of Selected Topics in Signal Processing*, *13*(2), 236–243. https://doi.org/10.1109/JSTSP.2019.2904212

Potsdam, E. (2022). Malagasy extraposition. *Natural Language & Linguistic Theory*, *40*(1), 195–237. https://doi.org/10.1007/s11049-021-09505-2

Qian, M., McLoughlin, I., Quo, W., & Dai, L. (2016). Mismatched training data enhancement for automatic recognition of children's speech using DNN-HMM. *10th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, 1–5. https://doi.org/10.1109/ISCSLP.2016.7918386

Qiu, L., & Benbasat, I. (2005). Online Consumer Trust and Live Help Interfaces: The Effects of Text-to-Speech Voice and Three-Dimensional Avatars. *International Journal of Human–Computer Interaction*, *19*(1), 75–94. https://doi.org/10.1207/s15327590ijhc1901_6

Ratianantitra, V. M., Razafindramintsa, J. L., Mahatody, T., & Manantsoa, V. (2019). Deep learning approach for Malagasy text summarization. *International Journal of Conceptions on Computing and Information Technology*, *7*(1), 12–17.

Řepová, B., Zábrodský, M., Plzák, J., Kalfert, D., Matoušek, J., & Betka, J. (2021). Text-to-speech synthesis as an alternative communication means after total laryngectomy. *Biomedical Papers of the the Faculty of Medicine of Palacký University*, *165*(2), 192–197. https://doi.org/10.5507/bp.2020.016

Ronquillo, C., & Currie, L. (2012). The digital divide: Trends in global mobile and broadband Internet access from 2000–2010. *NI 2012 : 11th International Congress on Nursing Informatics*, *2012*, 346. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3799113/

Saheer, L., Yamagishi, J., Garner, P. N., & Dines, J. (2014). Combining Vocal Tract Length Normalization With Hierarchical Linear Transformations. *IEEE Journal of Selected Topics in Signal Processing*, *8*(2), 262–272. https://doi.org/10.1109/JSTSP.2013.2295554

Serva, M., & Pasquini, M. (2021). Malagasy dialects in Mayotte. *EPL (Europhysics Letters)*, *133*(6), 68003. https://doi.org/10.1209/0295-5075/133/68003

Shen, J., Pang, R., Weiss, R. J., Schuster, M., Jaitly, N., Yang, Z., Chen, Z., Zhang, Y., Wang, Y., Skerry-Ryan, R. J., Saurous, R. A., Agiomyrgiannakis, Y., & Wu, Y. (2018). Natural TTS Synthesis by Conditioning WaveNet on Mel Spectrogram Predictions. *ArXiv:1712.05884 [Cs]*. http://arxiv.org/abs/1712.05884

Tabet, Y., & Boughazi, M. (2011). Speech synthesis techniques. A survey. *Proc. International Workshop on Systems, Signal Processing and Their Applications, WOSSPA*, 67–70. https://doi.org/10.1109/WOSSPA.2011.5931414

Tacotron 2 (without wavenet). (2022). [Jupyter Notebook]. NVIDIA Corporation. https://github.com/NVIDIA/tacotron2 (Original work published 2018)

Tuttle, K. R. (2020). Impact of the COVID-19 pandemic on clinical research. *Nature Reviews Nephrology*, *16*(10), 562–564. https://doi.org/10.1038/s41581-020-00336-9

Wang, Y., Skerry-Ryan, R. J., Stanton, D., Wu, Y., Weiss, R. J., Jaitly, N., Yang, Z., Xiao, Y., Chen, Z., Bengio, S., Le, Q., Agiomyrgiannakis, Y., Clark, R., & Saurous, R. A. (2017). Tacotron: Towards End-to-End Speech Synthesis. *ArXiv:1703.10135 [Cs]*. http://arxiv.org/abs/1703.10135

Wikipedia. (2022, April 14). *Wikipedia: The free encyclopedia*. Wikipedia, the Free Encyclopedia. https://en.wikipedia.org

Yang, F., Yang, S., Zhu, P., Yan, P., & Xie, L. (2019). Improving mandarin end-to-end speech synthesis by self-attention and learnable gaussian bias. *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 208–213. https://doi.org/10.1109/ASRU46091.2019.9003949

Zequeira Jiménez, R., Fernández Gallardo, L., & Möller, S. (2018). Outliers Detection vs. Control Questions to Ensure Reliable Results in Crowdsourcing.: A Speech Quality Assessment Case Study. *Companion Proceedings of the The Web Conference 2018*, 1127–1130. https://doi.org/10.1145/3184558.3191545

Zhang, C., Zhang, S., & Zhong, H. (2019). A Prosodic Mandarin Text-to-Speech System Based on Tacotron. *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 165–169. https://doi.org/10.1109/APSIPAASC47483.2019.9023283

Zhao, Y., & Hryniewicki, M. K. (2018). XGBOD: Improving Supervised Outlier Detection with Unsupervised Representation Learning. *2018 International Joint Conference on Neural Networks (IJCNN)*, 1–8. https://doi.org/10.1109/IJCNN.2018.8489605

Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., & He, Q. (2021). A Comprehensive Survey on Transfer Learning. *Proceedings of the IEEE*, *109*(1), 43–76. https://doi.org/10.1109/JPROC.2020.3004555

APPENDIX

Relevant default (English) parameters from the Tacotron 2 repository (*Tacotron 2 (without Wavenet)*, 2018/2022)

Experiment Parameters	
epochs	5000
iters_per_checkpoint	1000
dynamic_loss_scaling	True
ignore_layers	‘embedding.weight’
Data Parameters	
text_cleaners	‘english_cleaners’
Audio Parameters	
max_wav_value	32768.0
sampling_rate	22050
filter_length	1024
hop_length	256
win_length	1024
n_mel_channels	80
mel_fmin	0.0
mel_fmax	8000
Model Parameters	
n_symbols	len(symbols
symbols_embedding_dim	512
Encoder Parameters	
encoder_kernel_size	5
encoder_n_convolutions	3
encoder_embedding_dim	512
Decoder Parameters	
n_frames_per_step	1
decoder_rnn_dim	1024
prenet_dim	256
max_decoder_steps	1000
gate_threshold	0.5
p_attention_dropout	0.1

p_decoder_dropout	0.1
Attention Parameters	
attention_rnn_dim	1024
attention_dim	128
Location Layer Parameters	
attention_location_n_filters	32
attention_location_kernel_size	31
Mel-Post Processing Network Parameters	
postnet_embedding_dim	512
postnet_kernel_size	5
postnet_n_convolutions	5
Optimization Hyperparameters	
use_saved_learning_rate	False
learning_rate	0.001
weight_decay	0.000001
grad_clip_thresh	1.0
batch_size	4
mask_padding	True