university of groningen

campus fryslân

Age-controllable speech synthesis: A pilot study on English

Alice Vanni

university of groningen / campus fryslân

University of Groningen - Campus Fryslân

Age-controllable speech synthesis: A pilot study on English

Master's Thesis

To fulfill the requirements for the degree of
Master of Science in Voice Technology
at University of Groningen under the supervision of
Dr. T.P. Do (Voice Technology, University of Groningen)
with the second reader being

Alice Vanni (s5298873)

June 11, 2024

Acknowledgements

I would like to express my gratitude to my supervisor Dr Phat Do, for his support and the freedom he gave me in shaping this thesis. Thank you to Dr Vass Verkhodanova too, for the help with the Praat scripts and for the energy and passion she brings to the programme.

I also want to thank Voiseed's CTO Dr Lorenzo Tarantino for his precious pieces of advice in the early stages of this work.

I acknowledge the License Management of the University of Groningen and Campus Fryslân's Collection specialist for My Science Tutor corpus retrieval and the Center for Information Technology of the University of Groningen for their support and for providing access to the Hábrók high-performance computing cluster.

Thank you also to my fellow students who have always been supportive, in very practical ways too, and shared this journey with me.

Finally, I owe my deepest gratitude to my family, my partner Pietro and all my long-distance friends who encouraged my ideas and listened to my complaints for the whole year, making me feel loved and supported even with all these kilometres between us.

Abstract

This research attempts to implement age control in a text-to-speech (TTS) system to allow changing the perceived age of the synthetic voice while keeping the perceived speaker identity. The system uses a non-auto-regressive multi-speaker TTS model, namely FastSpeech2 (Chien et al., 2021) and was inspired by the pipeline outlined for ChildTTS (Jain et al., 2022). It uses Resemblyzer, a pre-trained speaker encoder, and entails an age encoder to extract embedding vectors used to generate speech by children, adults and elderly people. The system is developed for English using a corpus drawn from the Common Voice 17.0 English dataset (Ardila et al., 2020) and the My Science Tutor corpus (Pradhan et al., 2023). The model's performance was evaluated by acoustic analysis of the synthetic speech features and the calculation of Mel-Cepstral Distortion. The proposed system is designed to enhance the customisation of Speech Generating Devices (SGDs) and, additionally, to tackle the challenge of developing TTS systems for non-standard voices. The outcome of this research not only contributes to the broader understanding of voice personalisation techniques but also may play a part in providing new insight into the impact of the ageing process on voice. This will positively affect the industry, enabling more efficient creation of tailored voices, e.g. for Voice Assistants and vocal personas, as well as SGDs users. Age is an integral part of identity, and the ability to recreate a synthesised voice that a person identifies with can be an invaluable tool for those who have lost the ability to speak naturally.

Contents

1 Introduction

Part of who we are as people in society is influenced by our age, which can be understood not only as a biological fact but also as a part of our social identity (Johfre and Saperstein, 2023). Communicating with a voice that does not reflect our age can often mean communicating with a voice that does not match our identity. This case often applies to Speech Generating Devices (SGDs) users, who might be using these devices across different stages of their lives, spanning many years. The current advancement of technology makes speech synthesis systems capable of creating powerful tools and enabling everyone to speak with their own voice, expressing who they are through speech, too.

Nonetheless, SDGs often have only a limited set of voices and the level of customisation is low, especially in the long run. One of the implications is that people of different ages and in different life phases will most likely be using the same voice, with which they might not identify. To address the need for a match between the synthetic voice and one's age identity, I will research how to generate synthetic speech that can be customised according to the user's age throughout the years.

The state-of-the-art TTS engines are quite advanced, and cloning a voice from available data is not the challenge that was a few years ago (Hasanabadi, 2023). Nonetheless, voice cloning techniques do not solve the issue depicted above, for two main reasons. First of all, it is not given that people in need of an SGD can produce speech sufficiently good to clone their voice. Secondly, voice cloning will provide users with a voice that represents theirs at the moment of the recording. If the speech available for cloning is from a younger age or the person is in a phase of life that involves significant changes on a vocal level (e.g. teenage years), the cloned voice will soon be obsolete in terms of identity representation. For this reason, I believe the system I intend to implement can be beneficial, even in a moment in which TTS and voice cloning systems are very advanced[1].

There are already TTS models that address the issue of synthesising non-standard speech (i.e., speech not produced by a healthy adult male speaker). For example, Davatz et al. (2021) developed a vowel synthesiser for young, middle-aged and elderly adults, while ChildTTS, a model for children's speech synthesis, was successfully developed by Jain and colleagues (2022). On the contrary, to my current knowledge, no system was developed with the specific intention to synthesise a voice that can be parametrically shifted from old to young and vice versa, without having to re-train or fine-tune the model with specific data.

As it will be further discussed (2.1), speech from speakers of different age groups differs in many aspects. First of all, there are acoustic differences in the produced speech. This is due to changes in the vocal tract and in the respiratory system that occur with the aging process (Cho et al., 2021; Kuppusamy and Eswaran, 2022). Secondly, the morphosyntactic structure and the lexical choice are greatly affected by the age of the speakers (Cho et al., 2021). While the latter is quite self-evident to almost any language user, the former are more subtle and require a deeper analysis. This thesis will consider only the changes that occur on the phonetic level, the acoustic correlates of age in speech.

Non-standard speech, as I defined it above, can be considered a low-resource type of speech, since at the moment there is little availability of data that covers a wide age range. To be more precise, it is not the data themselves that are lacking, but rather the type of annotation required for the training. To be able to correctly learn the voice features of different ages the training data should

[1] I will leave to future research and developments the implementation of a system that better integrates voice cloning and age control.

contain information about the age of the speaker, for each training sample. This condition is often not satisfied by large corpora available. Big corpora often have this kind of metadata only for a subset of data, e.g. in the case of Mozilla Common Voice datasets (Ardila et al., 2020), while specific corpora with explicit age or age range annotation are even more scarce. As a consequence, I believe the development of this kind of model should put into practice methods developed for under-resourced languages.

This scenario unveils another shortcoming of current TTS systems. Although state-of-the-art TTS models demonstrate strong performances, they still face the challenge of using minimal training data while maintaining the same performances obtained with a larger dataset. This weakness adds up to the issues of controllability and customisation. This problematic point persists across various AI domains, where increasingly complex models require ever-larger datasets (Do et al., 2021). The growth in data requirements will lead to an increase in the number of languages and speaking communities for which we do not have enough data, making the case of developing solutions for low-resource languages (LRLs) more pressing than what it is now.

On a more general note, NLP tools and speech technologies are currently working and available for about 0.3% of the existing languages in the world (Magueresse et al., 2020). Given the overwhelming majority of languages not covered by today's technology, I believe that we should work to fill this gap. The speech tech community should channel its effort towards using fewer and fewer data so that we will not only be able to develop tools for LRL, but also for other kinds of under-represented voices.

For these same reasons, the age-controllable TTS system I am about to describe in this work is developed with having in mind LRL. As I will discuss in further detail throughout this thesis and especially in Section 6.1, the whole system was implemented to ease the fine-tuning and adaptation of it for LRL.

To sum up, this work aims to develop a text-to-speech model that will facilitate the customisation of voices for SGD users, enabling them to speak with a voice that is closer to the one they feel like their own.

Now that I introduced the reasons and general issues that motivated me to undertake this project, I will briefly outline the present thesis.

In the upcoming Section (1.1) I will detail my research question together with my hypothesis on how I achieved my objective and the expected outcome.

In Chapter 2 I will provide a short review of the state-of-the-art in three relevant fields, namely age-related vocal features (Section 2.1), age-related speech recognition and synthesis systems (Section 2.2).

Chapter 3 contains a description of the model (Section 3.1), the data (Section 3.2) and the training and evaluation methods used to reach my goal. The results and performance of such a system and the related discussion are covered in Chapter 4.

Before the final remarks (Chapter 6), in which I will cover some of the applications of the proposed work and its relevance, in Chapter 5 I will acknowledge and discuss some ethical concerns related to the presented system and more generally to TTS systems.

1.1 Research question and hypothesis

To tackle the issues introduced in the preceding paragraphs and fill the gaps in the current state-of-the-art, in the upcoming pages I will address the following research question:

> How can age-related vocal acoustic features be effectively parameterised and subsequently used in a TTS model to reflect different life stages in synthesised speech?

I hypothesise I can generate speech that accurately represents three age stages using a multi-speaker non-auto-regressive TTS model with age-parameterized training, namely *agingTTS*. The target age phases are childhood, adulthood and elder age. The hypothesis will be assessed through objective evaluation.

The falsification of this hypothesis contributes to highlighting which acoustic features of the human voice that correlate with age are not captured by machine learning models. As a consequence, it will provide support in delineating alternative methodologies for speech synthesis with age control. Some of the approaches that can be adopted to further ameliorate my system are already proposed in Section 6.1.

More precisely, building on the pipeline presented by Jain and colleagues (2022), chosen both for its low data requirements and for being suited for synthesising non-standard voices, I developed agingTTS, by modifying the multi-speaker implementation of FastSpeech2 (Ren et al., 2022) with learnable speaker embeddings by Chien et al. (2021)[2].

The target language for this pilot study is English, but the model has been developed having LRLs in mind, hence it is implemented considering the needs of this type of system.

All the relevant documents and code through which I fulfilled my research are available at https://github.com/AliceVanni/agingTTS.

[2] https://github.com/ming024/FastSpeech2

2 Literature review

The current chapter aims to provide an overview of the state-of-the-art in fields relevant to this work. What follows does not aim to be a fully comprehensive review of the current developments in these fields, but rather to give the reader the necessary elements to understand the process through which I arrived at the development of the Research Question stated in the previous Section.

The Chapter is structured as follows: first, in Section 2.1 I introduce some concepts about the influence of the aging process on our voice. This will equip the reader with the necessary knowledge to understand why it is possible to achieve age control over an artificially generated voice. I will then move on to describe current speech technology that relates in various ways to the age of the speaker, touching upon both Automatic Speech Recognition (ASR) systems and TTS models. This is the content of Section 2.2.

Before moving on to the proper literature review, let me describe the methods with which it was conducted.

Methods

For this literature review, I selected papers, articles and book chapters found via the Google Scholar search engine filtering the results with a time limit of 10 years (i.e., from 2014 on). Research done over 10 years ago was considered only if historically relevant for the field (e.g., landmark studies) and when relevant to provide theoretical background (e.g., illustrating a linguistic concept). Other selection criteria were:

- the language of the resource had to be English;
- the resource had to be freely accessible (no paywall);

The search was done using different keywords for the different components of this work. Only the titles relating to speech features were selected, even though the keywords used were sometimes more generic, to capture a wider range of titles. Finally, articles concerning pathological speech were not taken into consideration.

For the first topic (2.1), which investigates the age features in speech, the keywords used were mainly “age speech features”, “voice and age correlates” and “age effects on voice”. For the literature review on speech technology systems based on age (2.2), the keywords used were: “speaker age recognition”, “age-related ASR” and “voice age classification” for ASR systems and “age-related TTS”, “synthetic speech with age” and “age-parametric TTS” for TTS models. For the former search, the results mainly concerned both age and gender recognition systems for speaker identification and feature extraction, while very few relevant results were found for the latter. For this reason, the research was expanded by selecting papers and sources referenced in the previously selected resources, even though they did not contain the aforementioned keywords. In doing so, approaches involving Deep Neural Networks (DNN) were prioritised.

Table 1 below provides a summary of the literature discussed in the following paragraphs.

Table 1: List of references

Reference	Title	Research area
Linville, 1996	*The sound of senescence*	Age and Voice
Moyse, 2014	*Age Estimation from Faces and Voices: A Review*	
Huckvale and Webb, 2015	*A Comparison of Human and Machine Estimation of Speaker Age*	
Skoog Waller et al., 2015	*Can you hear my age? Influences of speech rate and speech spontaneity on estimation of speaker age*	
Skoog Waller and Eriksson, 2016	*Vocal Age Disguise: The Role of Fundamental Frequency and Speech Rate and Its Perceived Effects*	
Eichhorn et al., 2018	*Effects of Aging on Vocal Fundamental Frequency and Vowel Formants in Men and Women*	
Huff et al., 2020	*Can Computer-Generated Speech Have an Age?*	
Cho et al., 2021	*Lexical and Acoustic Characteristics of Young and Older Healthy Adults*	
Barkana and Zhou, 2015	*A new pitch-range based feature set for a speaker's age and gender classification*	Age and Gender ASR and Classification
Qawaqneh et al., 2017	*Deep neural network framework and transformed MFCCs for speaker's age and gender classification*	
Mei and Min, 2018	*Automatic Age Estimation Based on Vocal Cues and Deep Neural Network*	
Kuppusamy and Eswaran, 2022	*Convolutional and Deep Neural Networks based techniques for extracting the age-relevant features of the speaker*	
Yücesoy, 2023	*Speaker age and gender recognition using 1D and 2D convolutional neural networks*	
Luong et al., 2017	*Adapting and controlling DNN-based speech synthesis using input codes*	Age-related TTS
Davatz et al., 2021	*Source and Filter Acoustic Measures of Young, Middle-Aged and Elderly Adults for Application in Vowel Synthesis*	
Jain et al., 2022	*A Text-to-Speech Pipeline, Evaluation Methodology, and Initial Fine-Tuning Results for Child Speech Synthesis*	

2.1 Age and Voice

As already briefly mentioned in the previous Chapter, age is both a physical and a social matter (Johfre and Saperstein, 2023).

As a physical process, it involves the vocal tract as well as other elements influencing speech production (e.g. the respiratory system and muscle flexibility). Consequently, the acoustic features of a person's voice change with time, especially in terms of Fundamental frequency (F0) and Speaking Rate. This is confirmed by both acoustic and perception studies.

Speaking rate, together with F0, is pointed out as the main voice correlate of age by Skoog Waller and colleagues (2015). They showed that a speech rate manipulation of 10% already influences listeners' perception of the speaker's age. This is especially true when estimating older speakers' voices. The above study in fact showed that a slower speaking rate is more strongly associated with higher age. When the speaker is younger, and in a spontaneous speech context, the listener primarily relies on other cues, such as the lexicon, while speech rate is only a secondary cue. This conclusion is also confirmed by a further study by the same authors (Skoog Waller and Eriksson, 2016) which showed that only the speaking rate is used as a cue in age estimation by listeners, and the accuracy of estimating the real age of a speaker disguising their age by changing F0 and SR is pretty high.

Fundamental frequency also changes significantly with aging, but unlike the case of speaking rate, its direction of change depends strongly on gender. According to Linville (1996) and Eichhorn et al. (2018), the pattern of F0 in men goes from high in childhood, then lowers going into young adulthood to middle age and rises again into old age, due to physiological changes brought by aging (e.g. stiffness of vocal folds, changes in the larynx). Men's F0 peak in adult life is estimated at the age of 85. In women, the F0 remains quite constant until menopause, in which a lowering occurs, due to hormonal changes that have an impact on the vocal folds (Linville, 1996; Eichhorn et al., 2018).

In support of the conclusions outlined above, Moyse (2014) provided a review of studies investigating the ability of humans to recognize the age of the speaker. This review highlighted that age estimation from voices is more accurate for female voices than for male voices and that there is an influence on the length of the stimulus. It also highlighted that longer stimulus presentations yielded a better performance in age estimation. Nonetheless, according to Moyse (2014), it is not possible to estimate the exact age of a person only based on their voice.

Similarly to Moyse (2014), Huckvale and Webb (2015) compared the accuracy in age estimation between humans and machines by measuring the mean absolute error (MAE) of estimation. Based on their experiment, the highest performance of a machine was only 1.15 years more precise than the best human performance.

Given the different patterns of male and female voices over age and the difference in the estimation performance by both humans and computers, studies illustrate feature extraction and age classification ASR systems for both age and gender (see 2.2).

In this work, I will only consider the effect of the aging process on a phonetic level, but it is important to stress that, being age a social fact too, it also impacts other linguistic and sociolinguistic aspects.

First and foremost, the lexicon changes greatly across the lifespan. Children have a reduced lexicon since they are still acquiring language proficiency, while for young adults, adults and the elderly the lexicon varies due to social factors, i.e. cultural references, jargon, etc., but also in terms of diversity. Based on the results of Cho et al. (2021) on narrative speech in English[3], the elderly have a less diverse lexicon compared to younger people (more repetitions), while younger speakers tend to use more ambiguous words. Additionally, this study highlights differences in the use of morphological and syntactic structures: older speakers use shorter clauses and more inflected verbs compared to younger speakers (Cho et al., 2021). These aspects, despite being features of speech from different age groups, will not be considered in the present thesis.

As a final remark, given the topic of the present work, it is relevant to mention that age is not only perceived in human voices but in AI-generated voices too. Huff et al. (2020) showed that people perceive the (intended) age of a voice in artificially generated speech. Based on Skoog Waller and Eriksson (2016) findings, they also tested the possibility of manipulating such perception by changing the speed and the pitch of the speech. Their experiments used gTTS and WaveNet to generate the voices, and their outcome confirmed that age is perceived in AI-generated voices as well as in human ones.

2.2 Age-related Speech Technologies

In this section, I will provide an overview of the latest ASR and TTS systems that aim to classify and generate speech based on age, and consequently gender.

2.2.1 Age and Gender ASR and Classification

Age and gender speaker recognition is not a novel task (see e.g. Minematsu et al., 2002) and various approaches have been taken over the years, both in terms of Machine Learning architectures employed and in terms of the set of features used, which robustness has been considered a key issue.

In 2015, Barkana and Zhou (2015) presented a set of features based on pitch range (PR) for text-independent systems for age and gender classification. PR is selected as it is assumed to represent better than F0 the pitch variations of speech, and additionally, the authors argue that age correlates better with how rapidly the pitch changes over time. Such a set was tested on a Support Vector Machine (SVM) and K-Nearest Neighbors (kNN) classifier, and the performance was compared to different sets of features. The authors tested their hypothesis on the aGender corpus (Burkhardt et al., 2010) and showed that PR features yielded the highest accuracy. Especially in the case of age classification using kNN, the PR system achieved between 50% and 70% for the various classes, while the same architecture using MFFCs and Energy features led to around 25-40% accuracy.

In 2016, an age and gender classification framework using DNN and i-vectors was developed by Qawaqneh, Mallouh and Barkana (2017). Despite the findings of Barkana and Zhou (2015), the features used by Qawaqneh and colleagues (2017) are MFCCs transformed using a Bottleneck feature (BNF) extractor. The proposed work also used the aGender corpus for training and achieved higher

[3] The results are based on the Cookie Theft picture description task

overall accuracies by both BNF-I-vector and BNF-DNN classifiers (56.13%, 58.98%) compared to previous works on the same database. Nonetheless, the paper shows that the higher accuracy is again brought by the choice of the input features, not by the choice of the architecture. Transforming MFCCs with BNF gets rid of redundancies and additionally enables the caption of features in short-time utterances. Moreover, they retain the speaker's information and are language-independent.

Mei and Min (2018) work confirms the benefits of using a DNN-based approach for age estimation, as already shown in Qawaqneh et al. (2017). According to Mei and Min (2018), previous models did not perform well enough on age estimation mainly because of the lack of data available. The authors collected data from 128 Chinese and English speakers of ages ranging from 5 to 70 years old, dividing them into seven age groups for the training. The main contributions of this paper are the definition of relevant acoustic features for age estimation and the extension of the DNN-based approach for the task. The proposed algorithm follows these three stages:

> "Step One: Build an age group probability distribution at segment level from DNN. Step Two: Extract static features at utterance level. Step Three: Age estimation using Support Vector Regression (SVR)." (Mei and Min, 2018: 211)

Their results showed once again how the use of DNNs enhances the accuracy significantly compared to a GMM-based approach. Age estimation using Support Vector Machines (SVMs) yielded a 5.9 averaged error, while the averaged error for the GMM system using the Expectation Maximisation (EM) algorithm was 9.0.

The work done by Kuppusamy and Eswaran (2022) goes in the same direction illustrated above. Their work demonstrates again that the use of DNNs, especially with Convolutional layers (CNNs) yields the best performance in age classification tasks. In line with the work of Qawaqneh et al. (2017), they propose a feature extraction methodology that employs a CNN-DNN-based system with enhanced bottleneck features, which are considered noise-robust, and localized convolution filters, which have the function of normalizing the spectral variation brought by differences in vocal tract length of the speakers of different ages. The proposed approach was tested using four age categories and yielded the best results in combination with a GMM-SVM classifier, which had an 82% accuracy, while the same classifier with MFFCs and PLPs and with simple CNN features (i.e., without the addition of the bottleneck) had an accuracy of 59% and 77% respectively.

It has to be noted that while the first two works mentioned considered both age and gender, the succeeding two do not consider gender when extracting age features. This might be problematic, since gender- and age-related features are often overlapping, but it also simplifies the classification task, in that the distinctions to be made are less. Nonetheless, considering both the age and gender of the speaker might improve the accuracy of the recognition. An additional issue raised by Mei and Min (2018) and not accounted for in the following works is the cultural and language differences that might introduce additional variation.

Both the aforementioned shortcomings are considered in Yücesoy (2023) which tests two age and gender classifiers composed of 1D and 2D CNNs respectively. The input feature vectors used are made of 39 MFCC features and the delta and delta-delta of the first 13 MFCCs. The experiments use the Common Voice Turkish dataset, and the highest validation accuracy is obtained by the 2D model, which scored 94.40% accuracy. Surprisingly, the results reported in this paper are the opposite of what was shown by previous works. The above literature argues that the main improvement in age and gender recognition is brought by the choice of the input features, while the work of Yücesoy (2023) highlights the importance of the choice of the model for the classifier.

Contrary to previous literature, and in line with what Mei and Min (2018) concludes, Yücesoy (2023) also claims that even if age and gender features are not language-dependent, the developed system has to be considered language-dependent.

2.2.2 Age-related TTS

Unlike age and gender speaker recognition, TTS systems that aim to replicate voices of different ages are not widespread yet. There has been an effort towards children's speech synthesis, but to my knowledge, currently very few TTS models have been developed with the specific aim to synthesise speech that can be adjusted in relation to a target age.

One system that includes some degree of age control was developed by Luong and colleagues (2017). The aforementioned work aimed to develop a DNN-based speech synthesis system that is multispeaker, and is able to perform speaker adaptation with minimal effort and speaker morphing, that is: change the synthetic speech features with input codes. The latter is what interests us for the purpose of this work since the speech features to be modified include age and gender.

The TTS model was developed on Japanese, using native speakers and studio-quality data. The data were labelled per age of the speaker, which ranged from 10 to 89 years old. The authors chose to use age bands instead of the exact, raw number and divided the speakers into ten age bands, obtaining roughly eight speakers per group.

The findings concerning age manipulation are ambivalent. As reported by the authors, the differences between the synthesised voices with different age codes are present and audible. Nonetheless, this is true only for distant age groups. Additionally, the degree of perceived differences between the different age groups was not specifically tested.

Additional work related to the synthesis of speech from different age groups was done by Davatz et al. (2021) on Brazilian Portuguese. Specifically, this prospective study focused on the realisation and further analysis of vowel synthesis, how speech characteristics change from young to elderly adults, and how this information can be used for synthesising speech in the target age. The authors investigated the application of source and filter acoustic measurements to vowel synthesis for three adult age groups, defined as 'young' (18 to 45 years old), 'middle-aged' (46 to 60 years old) and 'elderly' (61 to 80 years old).

The data were obtained by recording 162 adult native speakers of Brazilian Portuguese, with no known speech impairments. The collected speech consisted of a sustained production of vowel /a/, recorded in a sound-proofed studio. These audio recordings were then analysed using LPC and they extracted the values of the fundamental frequency, the first four formants and four bandwidths. The extracted values were then employed to perform vowel synthesis, which was in turn analysed to highlight the most significant features of each age group speech.

The findings from Davatz and colleagues (2021) align with previous studies, highlighting a higher F0 for young and middle-aged women compared to elderly women. Nonetheless, no relevant acoustic measurement was found to discern the three age groups across male participants.

However, this study does not systematically evaluate the perception of the generated voices by subjective evaluation. The conclusions were only drawn by the acoustic analysis of the synthesised audio samples. Furthermore, even though this study provides insights into the differences in speech through the lifespan, it only attempted to synthesise static vowels, with no linguistic content.

As already mentioned, a significant work for the present project is the one described in Jain et al. (2022). Even though this work solely focuses on the synthesis of children's speech, I believe the pipeline outlined can be effectively applied for synthesising various types of age-related speech and other under-resourced voice types, given that the system was developed to perform good with a low amount of data. The paper describes the implementation of ChildTTS[4], a multispeaker TTS system for children's speech. Such a system has three modules:

- A speaker verification phase, which employs a generalised end-to-end model;
- An acoustic model based on a modified version of Tacotron 1 (Jia et al., 2019);
- A vocoder waveform generator.

The first step generates the speaker embeddings, which represent each speaker; the second step is the core of this TTS system and it is trained using the speaker embeddings generated in the previous stage. Finally, there is the waveform generator, a WaveRNN with Gated Recurrent Units (GRU), that generates the output speech.

The training of ChildTTS is done using first adult speech samples and then fine-tuning the whole model using children's speech. The data for fine-tuning was drawn by selecting a subset of the My Science Tutor (MyST) corpus (Ward et al., 2021; Pradhan et al., 2023), called TinyMyST.

The goodness of this system was measured by developing a specific set of evaluation criteria for children's speech, which cannot be measured with the same criteria that apply to adult speech. The results showed a close MOS score for the synthetic child speech generated with ChildTTS and the ground truth, highlighting the good performance of the model.

All the approaches described above, however, have failed to address the larger question of whether and how it is possible to develop a TTS system that can take age as an input parameter to synthesise speech. Filling this gap in current research is the aim of this thesis.

2.3 Conclusions

In this chapter, I have explored a range of studies focusing on age-related vocal features and their use in voice technology systems, with an emphasis on advancements in speech synthesis technologies. Notably, while there is progress in recognising speech with age-specific attributes, a gap persists in the synthesis technologies. Furthermore, the literature underscores the need for TTS systems that can dynamically adjust to represent various age-related vocal characteristics without extensive retraining or fine-tuning for each age group.

The synthesis of age features in voice remains under-explored. Existing research focuses on a narrow age range, such as children's speech, often neglecting the nuanced spectrum of vocal changes throughout the entire lifespan.

[4] https://github.com/C3Imaging/ChildTTS

This backdrop of technological advancement juxtaposed with unmet needs in TTS with age control directs us to the research question anticipated in the previous chapter. In the next chapter, I will present agingTTS, the system I developed to address the gaps highlighted above and find an answer to my research question.

3 Methodology

To answer my research question on how to parametrise age-related acoustic features of the human voice to use them in a TTS system, I started from the pipeline used for the implementation of ChildTTS (Jain et al., 2022) illustrated in Section 2.2. As previously mentioned, I employed a multi-speaker implementation of FastSpeech2 with learnable speaker embeddings (Chien et al., 2021) to train agingTTS on English.

The implementation described in this chapter was realised using multiple datasets of English speech, given that the availability of the necessary data was scarce. The choice of working on English was driven by the high availability of data in this language, which compensates for the low amount of data with age information. For the training of the English model, I employed Common Voice English in its latest version (17.0; Ardila et al., 2020) to gather adult and senior speech data. To these, I added a subset of data from the My Science Tutor corpus (Pradhan et al., 2023), a collection of children's speech interacting with a virtual assistant (see Section 3.2).

To investigate the performance of the model and compare it with the literature on age and voice, I analysed the acoustic features of synthesised output.

In this chapter, I will detail the procedures outlined above in the following order:

- Section 3.1 outlines the details of my implementation of the TTS model;
- Section 3.2 describes the features of the dataset I built to achieve my goal;
- Section 3.3 displays the experimental setup discussed in the following chapters;
- Section 3.4 illustrates the evaluation methods adopted to verify the hypothesis;

3.1 Model description

My work originates from the system developed by Jain et al. (2022). As described previously, the authors successfully developed a TTS model to synthesise children's speech, and they achieved this result by using a speaker encoder, which was first trained with data from adult speakers and then fine-tuned with children's speech data from a reduced version of the corpus I also adopted, the MyST corpus (see Section 3.2.2). Their pipeline further entailed the use of Tacotron2, an autoregressive model for TTS, and WaveRNN as a vocoder to synthesise the waveform.

In developing agingTTS, I adopted the same pipeline outlined above, but I made some changes to the architecture by integrating it with elements based on the conclusions drawn by Do et al. (2023b) for the choice of the acoustic model and the vocoder.

Similar to ChildTTS (Jain et al., 2022), agingTTS includes three main elements:

- An encoder module;
- An acoustic model;
- A waveform generator.

Even though studies showed that Tacotron2 can achieve better performance in low-resource environments compared to FastSpeech2 and also Deep Voice 3 (Gopalakrishnan et al., 2022), as already anticipated in the introductory section (1), the model I adopted for the realisation of the present system is FastSpeech2 (Ren et al., 2022). This choice was driven by multiple factors.

First of all, the higher training speed of FastSpeech2 made it a more suitable candidate for the time given to complete the present work, as well as the reduced inference time compared to Tacotron2 (Ren et al., 2022). An additional reason can be found in the overall higher performances achieved by this model compared to Tacotron2, thanks to the variance adaptor module (Ren et al., 2022). Being FastSpeech2 a more advanced, faster and better performing model compared to the one used in Jain et al. (2022), I chose to work with it.

FastSpeech2 was also judged particularly suited for the realisation of a TTS system with age control since it is already shown empirically to work well with a trainable speaker encoder (Chien et al., 2021), together with the variance adaptor. These two elements were exploited to implement the age encoder, which will be described in Section 3.1.2, and the age control during inference. Nonetheless, as it will be discussed in Section 3.1.1, the original implementation of the speaker embedding layer was substituted by a GE2E voice encoder trained separately (see Section 3.1.1). This separate speaker encoder and the age embedding layer are the main innovation elements of my implementation.

Moreover, the choice of this architecture follows the conclusion of Wells and Richmond (2021) and the subsequent implementation described in Do et al. (2023b). In both these works FastSpeech2 has been shown to perform better in combination with articulatory features, which have been highlighted as the best input features for LRLs TTS. Even though the implementation of such features was not possible in the present study, it is the next step to take to enhance this system (see also 6.1) and the use of FastSpeech2 paves the way for this.

Finally, by using FastSpeech2, the waveform generator is not only integrated into the system but also the RNN vocoder used in ChildTTS is replaced by a GAN-based vocoder, namely HiFi-GAN. This takes up the recommendation made by Jain et al. (2022) in their final remarks and follows the methodology applied by Do et al. (2023b), which is relevant for the same reasons laid out above. For these reasons, agingTTS uses the pretrained HiFi-GAN vocoder available in Chien et al. (2021)'s repository.

While the reasons to choose FastSpeech2 were many, there was one major disadvantage in its original implementation. The original structure as outlined by Ren et al. (2022) in fact did not support multi-speaker training, which was necessary to achieve my goal. To train a single-speaker TTS model, I would have needed speech from the same person in different phases of their life, and even though this is not completely impossible, this kind of dataset is definitely harder to obtain rather than a multi-speaker dataset with various speakers. This had two consequences. The first relates to the choice of the multi-speaker implementation of FastSpeech2, which fell on the one presented in Chien et al. (2021)[5]. Secondly, the need for multi-speaker data had an impact on the choice of the training data, together with other factors that will be further investigated in Section 3.2.

In conclusion, the implementation of agingTTS presented in this thesis includes only two substantial changes in the architecture of the multi-speaker FastSpeech2 model by Chien et al. (2021). The first and foremost addition to the model is the age embedding layer, which will be detailed in Section 3.1.2. The second addition to the original model is the GE2E model in place of the embedding layer to model speakers. This will be the focus of the upcoming section.

[5] The code is available at https://github.com/ming024/FastSpeech2 under MIT License

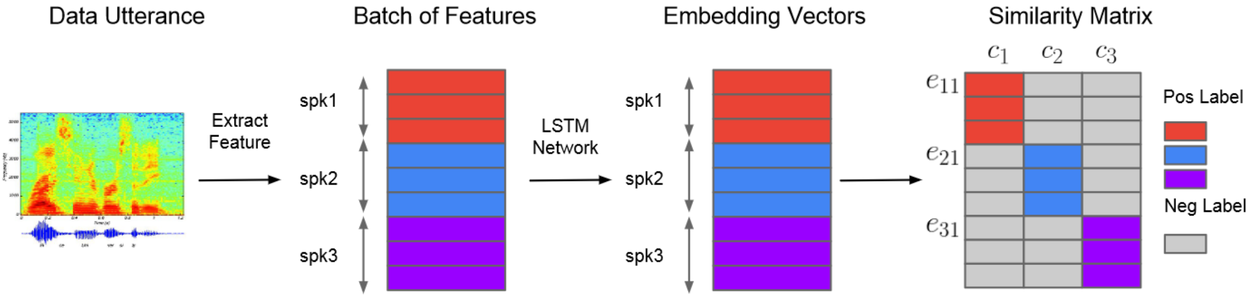


Figure 1: Overview of the system by Wan et al. (2020). The different colours signal different speakers.

The full implementation of agingTTS, together with the details about its use for both training and inference is available on GitHub under MIT License at the following URL:
https://github.com/AliceVanni/agingTTS.

3.1.1 Speaker encoder

As mentioned earlier, the multi-speaker FastSpeech2 implementation that I adopted as a basis for my model included a speaker encoder. The native speaker encoder was made of an embedding layer with 256-dimensional speaker embeddings and it was jointly trained with the rest of the architecture. To improve the quality of the speakers' representation, I replaced the native speaker embedding layer with a Generalized End-to-End (GE2E) model.

Such GE2E model was originally developed by Wan et al. (2020) for speaker verification. The system is composed of a Long-Short Term Memory (LSTM) network with a final linear layer (see Figure 1). The LTSM takes in input a batch of vectors of speech features x_{ji} extracted from each utterance i of speaker j. The batch is input at once and it is composed of N speakers with M utterances. The embedding vector e_{ji} is then calculated as the L2 normalisation of the network output $f(x_{ji};w)$, where w are the weights learnt by the LSTM network.

$$e_{ij} = \frac{f(x_{ji},w)}{||f(x_{ji},w)||_2}$$

Finally, for each speaker, the similarity matrix $S_{ji;k}$ is computed, which gives the degree of similarity between the input speakers. The similarity matrix $S_{ji;k}$ is calculated as the cosine similarity between each speaker embedding e_{ji} and all the centroids c_k.

For the present thesis, I adopted the implementation of Wan et al. (2020)'s work developed by resemble.ai, called resemblyzer[6]. The code is openly available for use under the Apache License.

3.1.2 Age encoder

My final goal is to enable the users to specify how old the voice they generate should sound. This is achieved by adding age information to the synthesised speech, in order to change the speech features towards the target age. There are two possible implementation approaches, the first involves the explicit indication of the target age, and the second implies the creation of age groups. The latter

[6] GitHub repository: https://github.com/resemble-ai/Resemblyzer; Python package documentation: https://pypi.org/project/Resemblyzer/

entails the *a priori* selection of age groups, which is to some degree an arbitrary decision, given that, as it will be argued in more detail below, there is no fixed and unanimous way to divide age into categories. However, as already mentioned in Section 2.1 humans do not have the ability to estimate the exact age of a speaker (Moyse, 2014). For this reason, I decided to work with a categorical representation of age. This *a priori* decision might have had consequences on the performance of the systems, nonetheless, it is also driven by the need to identify categories that could easily be modelled by the age encoder.

The selected age groups are:

- ‘child’ from 7 to 11 years old,
- ‘adult’ from 20 to 49 years old
- ‘senior’, which includes over-70 speakers.

The exclusion of the teens is based on the fact that in the 12-19 age range too many changes happen in a person’s voice, due to the hormonal and physical changes that adolescence brings. The 60-69 age range was excluded for two reasons. Firstly, a practical reason: to ensure a higher detachment between the ‘adult’ and the ‘senior’ categories, this age band was discarded. Moreover, there is no agreement on the starting age of seniority. According to the World Health Organisation (WHO, 2022), a person can be considered elderly from 60 years old onwards, nonetheless, unlike in the definition of adolescence (WHO, 2019), no biological or physiological event marks the beginning of this life phase. Elderly age is in fact sometimes considered to start at 60, but sometimes at 75 years old (e.g., Ouchi et al., 2017). Hence, excluding the 60-69 age range addressed both a practical and theoretical need.

With the data described above, I trained two different age encoders to extract age embeddings for ‘child’, ‘adult’ and ‘senior’ categories. The two age encoders are both based on the speaker encoder as found in Chien et al. (2021)’s implementation, but they differ in the dimension of the hidden layer.

To test what is the best configuration to extract age features using an embedding layer, I trained two models. First I used 256-dimensional speaker embeddings for the age embeddings, then I tested the efficacy of introducing a bottleneck in the age encoder by reducing the embedding dimension to 128 units. The bottleneck output is then input into a sequential projection layer that expands it to a higher dimensionality to a 256-dimensional vector. This layer comprises two fully connected linear layers separated by a Rectified Linear Unit (ReLU) activation function that introduces non-linearity.

In both models, the embedding layer is a simple lookup table that outputs 256- and 128-dimensional continuous vectors representing the discrete age categories in the latent space. The output is the age embeddings for the input categories. The age embeddings are initialized from scratch and trained together with the rest of the architecture.

The result of this age encoder is finally added to the encoder’s hidden sequence, as for the original speaker embeddings.

Even though I worked with three age groups, as explained at the beginning of this section this is a design choice, not a hard requirement for the development of the system. For this reason, the age encoder is already predisposed to generate age representations for more and different age groups. The flexibility of the age encoder will facilitate the development of a system that can model age with a higher accuracy through the use of more fine-grained age groups, or even the exact age of the speakers.

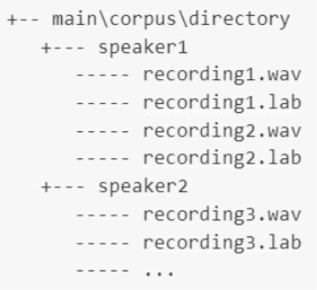
3.2 Datasets: description and preparation

The training and testing of the model described in the previous paragraph was done using parts of two datasets of English speech.Before diving into the description of each one of them, I want to specify that the corpora used for the development of my model were not originally collected with a TTS task in mind. Both English speech datasets were collected for ASR purposes and all of them are crowd-sourced, hence the quality of the data is not as high as it usually is for TTS systems. The consequences of this will be discussed further in Section 6.1. The meta-data requirements of the present system left no other choice than the use of the datasets I am about to present in the following sections.

All of the selected corpora went through a preparation procedure that involved mainly the meta-data files that came with them. Since the different datasets had different metadata formats, the first step was to make them coherent and homogeneous, to facilitate the further processing of the data. The template used and applied to every corpus's metadata document is an edited version of the metadata provided together with the Common Voice 17.0 English data, the 'validates.tsv' file to be precise. Starting from this file, only the columns with information relevant to my purpose were maintained, namely the speaker identification code, the name of the corresponding audio file, its transcription, gender and age of the speaker. The 'accents' column was also kept in case the presence of multiple English varieties turned out to be problematic. In addition, I included duration information for each file, which was retrieved either from a separate file, when available or directly from the audio clips.

The sampling rate of the audio files was also adjusted to make it homogeneous. The lowest sampling rate was found in the MyST data, and it was 16000 Hz, while the highest was found in the CV17 data, and it was 48000 Hz. All the audio files were resampled at 22050 Hz to match the sampling rate of the pretrained HiFi-GAN vocoder.

Finally, I restructured all the datasets in directories following the structure used by Prosodylab-aligner, which is also the style adopted by my reference implementation of FastSpeech2 (see Fig. 2). The resulting sub-directories were then merged into one in order to have all the data in the same place.



```
+-- main\corpus\directory
   +--- speaker1
      ----- recording1.wav
      ----- recording1.lab
      ----- recording2.wav
      ----- recording2.lab
   +--- speaker2
      ----- recording3.wav
      ----- recording3.lab
      ----- ...
```

Figure 2: Directory structure of *agingTTS* dataset

The following sections describe the CommonVoice English 17.0 dataset and the MyST corpus respectively, and the subset selected for the training of the present model. From now on, I will refer to the resulting dataset as the agingTTS dataset.

Every section has the same structure:

- Overview of the corpus;
- Reason of choice;
- Specific preprocessing needs;
- Final dimensions of the dataset in analysis.

3.2.1 Common Voice English 17.0

Common Voice Ardila et al., 2020 is a project by the Mozilla Foundation that collects open-access datasets in various languages. The collection of all the datasets is community-driven and the data are crowd-sourced. The whole process of collection is bottom-up and easily accessible to everyone, from the recording of speech to the validation of the data.

The corpus is not unproblematic, since the validation is not further checked by any expert and the process is done by many individuals, the outcome might not be as clean as when done by a single person with expertise. Additionally, the validation is done by people who self-declare native speakers, but there is no way to verify it.

Despite this, Common Voice currently represents one of the biggest speech collections freely available, and this represents a non-negligible advantage over other high-quality corpus. Additionally, the recordings come from a wide range of people, not only in terms of languages and accents but also in terms of demographic distribution.

This variety of demographics, together with the ease of accessibility to the data, was the main reason for which this corpus was chosen. The speaker's metadata collected prior to the audio recording includes age and gender information, which are core pieces of information for the present study. Unfortunately, the collection of such speaker's demographic data is not mandatory, as a consequence, not all the speakers provide such data.

To obtain the higher amount of data with age and gender information, the latest and biggest release of Common Voice English was chosen, version 17.0, released on March 20th, 2024[7].

Common Voice English 17.0, hereafter referred to as CV 17, comprehends 3508 hours of recorded speech, of which 2615 hours have been validated, from over 90000 speakers. Of the total recordings, less than 40% have no information about the age and gender of the speaker. For this reason, these data were discarded, together with the unvalidated ones. As already mentioned in 3.1.2, CV 17 was used only for the collection of data from 'adult' and 'senior' age bands, which means data from speakers under the age of 20, which represents 6% of the total data, were not included.

This selection was further pruned by the selection of a balanced of samples, based on the duration of utterances from age and gender subgroups of the corpus. The lowest amount of data was found in the 'senior' 'female' group, and it counted around 6 hours of speech. This data slice sets the default dimension of each other subgroup. This balancing resulted in a dataset of roughly 24 hours of speech by 15332 speakers (12 hours per age group).

I will be referred to the resulting, filtered CV 17 subset as FilteredCV17.

3.2.2 MyST - Children's speech corpus

My Science Tutor Children's Conversational Speech corpus (Ward et al., 2021; Pradhan et al., 2023), referred to as MyST corpus, is a collection of spoken data from children from the U.S.A., attending the 3rd, 4th and 5th grade (elementary school), ranging from 7 to 11 years-old. The corpus was collected as part of the My Science Tutor project (Ward et al., 2011) which aimed to improve science learning in elementary school children through short and conversational lessons with a virtual tutor, Marni. As reported in Ward et al. (2011), the dialogues are about 15-20 minutes long and their focus in on the student's ability to express themself. This is achieved through open-ended questions by the virtual tutor, to which the child is encouraged to think and explain their thoughts autonomously.

[7] Downloaded on 15/04/2024, link to the data: https://commonvoice.mozilla.org/en/datasets

The data gathered from these interactions consist of over 470 hours of conversational speech from 1 371 elementary school students, recorded from 2008 to 2017. The data collection involved two phases. The first stage had the goal of collecting data from the widest variety of students possible, while the second phase aimed for complete coverage of the material from single students. The data obtained from the first phase cover four topics, which correspond to four teaching module from all three grades. The topics are “Magnetism and Electricity”, “Mixtures and Solutions”, “Variable” and “Water”, for a total amount of 421 students who produced 109 hours of speech, all transcribed. In the second phase, only children attending the last two years of elementary school were included, for a total of 950 students. The total number of hours recorded in this stage was 364, of which only 115 were transcribed. The teaching modules covered were “Energy and Electromagnetism”, “Mixtures”, “Sun, Moon and Planets, “Soil, Rocks and Landforms”, and “Living Systems”.

As clearly stated in Pradhan et al. (2023) the corpus was created to improve ASR for children, and also to improve AI-driven education approaches. This implies that, even though data cleanup and preprocessing procedures have been applied to the raw data, the audio quality is not as high as usually desired for TTS applications.

Another critical point of the corpus is the lack of variety in the topics of the utterances. The fact that there is a restricted set of subjects of conversation entails a restricted kind of lexicon, and being the subjects of scientific nature, the lexicon is also quite unusual for an 8-year-old, e.g., “the citric acid goes in that 50 epsom salt and the other 15”. Nonetheless, since the data are conversational, there are also a number of occurrences of everyday phrases, such as “how are you”, “I am tired”, “we talked about...” or “we’ve been learning about ...”.

Despite the above, the MyST corpus has been proven effective in the training of ChildTTS (Jain et al., 2022). Additionally, as for the Common Voice dataset, this corpus is freely accessible and open source for academic purposes. These two are the reasons why I selected it for the training of my systems.

As pointed out above, not all the audio samples have the corresponding transcription. Since the transcription of children’s speech with automatic tools is not reliable, and the data missing the transcription are more than 50% of the full dataset, I discarded all the data without a transcription available. Audio samples with inaudible or unclear speech were also excluded from the training data. Such data were identified based on their transcription, which had annotation for non-verbal sounds too (e.g. noise, side speech from adults or the virtual assistant).

To use the corpus, I restructured the directory as described at the beginning of this Section and created the metadata files based on CV 17’s.

The resulting dataset was further cut to extract a subprotion with a duration comparable with that of the FilteredCV17 corpus. To do so, I extracted a maximum of 20 utterances for each speaker until it reached the duration of 12 hours. This subset of the MyST corpus, which I will call FilteredMyST, comprehends 5390 utterances from 281 speakers.

I will refer to the resulting dataset, composed of the combination of FilteredCV17 and FilteredMyST, ag *agingTTS dataset* and it comprises 36 hours of speech uttered by 3768 speakers of different varieties of English. The audio files have various durations, ranging from 2 to 10 seconds.

3.3 Experiments

As anticipated in 3.1.2, I trained two agingTTS models that differ only in the structure of the age encoder. The baseline model, which from now on I will simply refer to as *agingTTS*, entails an age encoder with 256 hidden units. The second model comprises a bottleneck in the age encoder. From now on, I will refer to this second model as *agingTTS-BN*. The training of both models was achieved using the agingTTS dataset.

The data were aligned with their transcription using the Montreal Forced Aligner (MFA; McAuliffe et al., 2017) with the pretrained language, acoustic and G2P models[8]. Through MFA, I also obtained the ground-truth durations of each training sample required by the Variance Adaptor, while pitch and energy values were extracted at the phoneme level, employing the code provided by the source multi-speaker FastSpeech2 implementation, with some slight adjustments.

While the data were undergoing the preprocessing I just described, I used the pretrained Resemblyzer to extract the speaker embeddings from those very same data, using the audio files only.

Both models were trained with the Adam optimiser (betas: 0.9, 0.98, epsilon: 0.000000001) with a batch size of 10 and a learning rate adjusted to a combination of warm-up and annealing steps, which had values of 4000 and 300000, 400000, 500000 respectively, with an annealing rate of 0.3. No weight decay was applied, while the gradient clip threshold was set to 1.0. The models were trained for a total of 100000 steps.

The training was not brought further due to time constraints, nonetheless, FastSpeech2 models can already be successfully used after the first 10000 training steps. This is clearly shown by the TensorBoards of the present and other models (e.g. Chien et al., 2021), from which it can be observed that the main improvements in the loss curves happen in the first 10000 steps.

The final models' checkpoints, together with more details about the implementation and training of agingTTS, can be found on the already mentioned GitHub repository[9].

The checkpoint was used to synthesise 18 sentences from 8 speakers, selected from the agingTTS dataset. This allowed me to make precise comparisons between the synthesised speech and the Ground Truth (GT). The list of the selected sentences with their corresponding speaker can be found in Table 2.

Table 2: List of sample sentences with the corresponding speaker's age group

Filename	Sentence transcription	Age group
12225c_01_gt	Uhm I'm not really sure	Child
12225c_02_gt	It rises from the east and sets in the west	Child
13027c_03_gt	It's about giving light to our earth and room	Child
13027c_04_gt	Maybe it means that it's the pathway of the light bulb	Child
13057c_05_gt	They make the energy flow	Child
13057c_06_gt	That it's connected in the right places	Child
3835a_07_gt	We are rolling without keys right now.	Adult
3835a_08_gt	Despite his lack of free time, he was able to continue writing.	Adult

[8] Available here: https://mfa-models.readthedocs.io/

[9] https://github.com/AliceVanni/agingTTS

Filename	Sentence transcription	Age group
6211a_09_gt	You can lead a horse to water, but you can't make him drink.	Adult
6211a_10_gt	What do you say?	Adult
6345a_11_gt	This turned into pleurisy complicated by pericarditis.	Adult
6345a_12_gt	Fares is married to Hala Fares.	Adult
16135s_13_gt	That's some hat you got on there.	Senior
16135s_14_gt	Traveling alone is good for meeting new people.	Senior
16135s_15_gt	Her manipulation failed	Senior
17615s_16_gt	He was buried in the Pantheon.	Senior
17615s_17_gt	The connection, she claims, is purely coincidental.	Senior
17615s_18_gt	He has also been a member of many other short-lived bands.	Senior

Each of these sentences was synthesised with *agingTTS* (the baseline model) and *agingTTS-BN* (the model with the bottleneck), applying all three age control groups. This led to a total of 108 samples.

The speech samples generated with *agingTTS* and *agingTTS-BN* have been evaluated and analysed according to the procedure outlined in the following section.

3.4 Performance analysis

The output audio of the two architectures was analysed by two means. I conducted an acoustic analysis of the synthesised speech to verify whether the output with different age controls differed on the features reported by the literature as relevant for different age groups, namely F0 and speech rate. Additionally, I calculated the Mel-Cepstral Distortion coefficient to measure the degree of difference between the synthetic and natural speech. This aimed to measure how close the synthesised speech of the two models is to the natural one. This comparison will also allow me to define which of the two models performs better.

The acoustic features of the output speech were analysed using Praat (Boersma and Weenink, 2023) and their statistical significance was tested by applying a one-way ANOVA test in R (R Core Team, 2021).

Given the results reported in the literature (2.1), I compared the mean F0 of child, adult and senior synthetic speech and their differences in speech rate.

The pitch listing from which the mean and range of F0 were extracted was generated using a Praat script derived from Lennes (2003) repository. The specific script can be found in my GitHub repository. The speech rate was calculated by considering the number of syllables per second. Similar to the pitch listing, the speech rate values were extracted using the Praat script by Nivja H. de Jong and Heeren (2021)[10].

Such analysis aims to validate the modelling of these features by the age encoder of the two experiments.

In the upcoming chapters, I will report and discuss the results of this evaluation.

[10] Both Praat scripts can be found here: https://github.com/AliceVanni/agingTTS/acoustic$_{a}nalysis$

4 Results and Discussion

The present Chapter illustrates the outcome of my work. The work included two experiments that tested whether age controllability in a TTS system could be achieved with a non-auto-regressive architecture, and which is the better-performing architecture. Both models employed an age encoder in their structure that allowed the modelling of age features of speech through embeddings. As previously explained, the difference between the two models tested lay in how these age embeddings were extracted. In the baseline model *agingTTS*, the age encoder extracted 256-dimensional vectors, having the same size as the speaker embeddings. The model with bottleneck, *agingTTS-BN*, instead entails an age encoder that modelled the age features first as a vector of 128 dimensions, then these vectors are projected into a higher-dimensional space.

Moreover, to evaluate in more detail the output audio and check for the ability of the architectures to model the age-related characteristics of the voice, an acoustic analysis was conducted, together with Mel-Cepstral Distortion analysis. The acoustic analysis aimed to verify if the features highlighted by the literature as the core elements of distinction of speech of different age groups were effectively caught and modelled by the base and bottleneck models. The details and outcome of such inspection are reported in Section 4.1.

4.1 Acoustic analysis

The acoustic analysis of the synthesised samples and GT speech was achieved through Praat, while the statistical analysis of the outcome was done using R.

This analysis compared the GT features with the features of the audio synthesised using *agingTTS* and *agingTTS-BN*. The acoustic features considered were the pitch and the speaking rate, as these two are pointed out by the literature as the main correlates of age in the voice.

4.1.1 Pitch

As already discussed in Section 2, the fundamental frequency, or pitch, of speech is one of the most easily perceivable dimensions of variation between speakers from different age groups. According to previous studies (see 2) children have the highest mean pitch among all the age groups, while adult speakers generally have the lowest since there is an increase of F0 in the latest phase of life.

To verify whether the models had caught these age-related features of the human voice, I used a Praat script based on Mietta Lennes' script for getting the pitch maximum[11] to extract the pitch listing from GT and synthesised audio. To better visualise the differences across the scenarios, I plotted the pitch listing using Python to obtain the pitch contour of each sentence. The pitch and time stamps were normalised per speaker and then plotted following the same criterion.

The pitch contour highlights that there is indeed a difference in the pitch with which the different age-controlled sentences were synthesised. Looking at the plots, the children's synthesised sentences (symbolised by pink lines) are realised with a higher pitch compared to both adult and senior's speech (orange and blue lines respectively) for all speakers. This can also be very clearly perceived by simply listening to the synthesised audio samples[12]. Unlike the findings reported in 2, the senior and adult synthesised speech do not present an evident difference in pitch contour, even though these

[11] Github repository: https://github.com/lennes/spect

[12] The reader can listen to these samples at https://alicevanni.github.io/agingTTS/

sentences are not synthesised with exactly the same contour. It is also quite self-evident that the GT pitch contours (i.e., the black lines) do not correspond to the synthesised speech from the same age group for children's speech. As it can be observed in the upper plot in Figure 3, the GT children's speech is realised with a lower pitch compared to the synthesised speech from the same speaker.

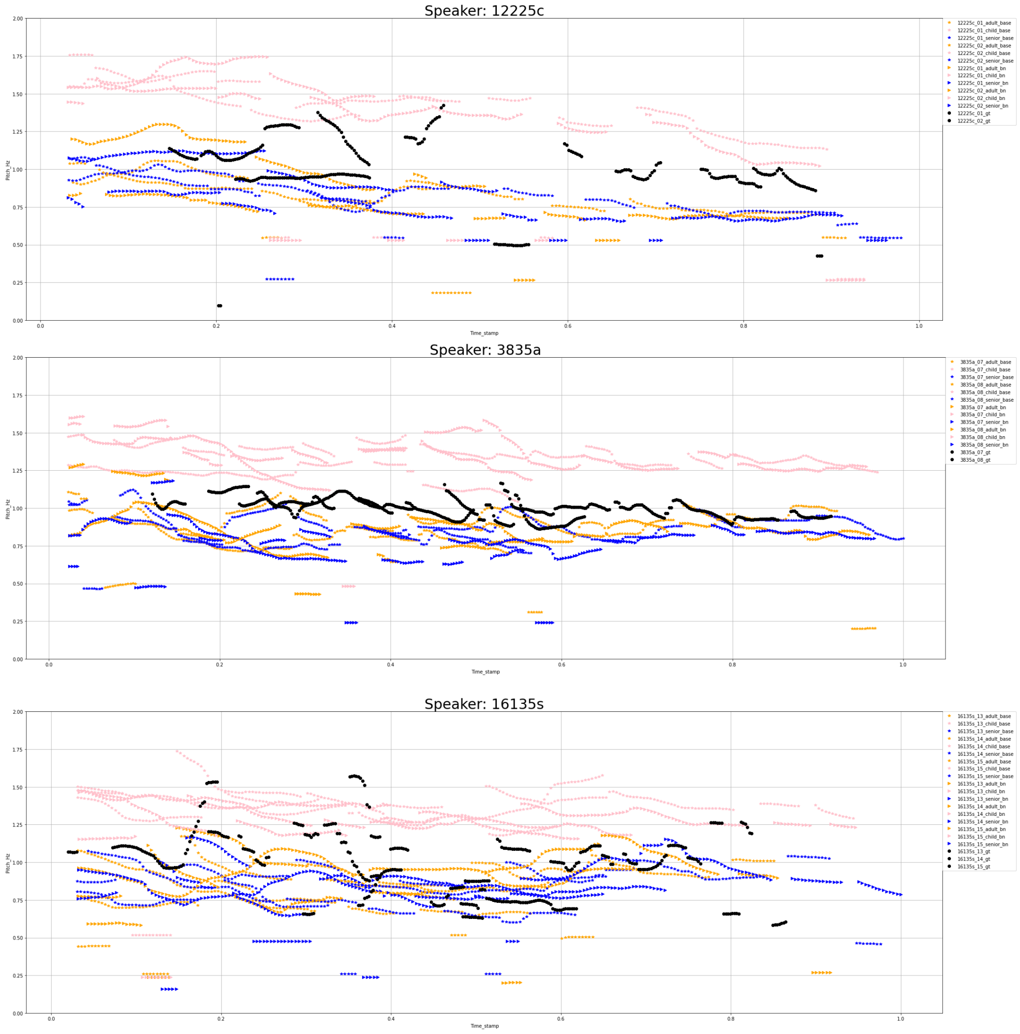


Figure 3: Aggregated pitch contour of sentences synthesised with the baseline (*agingTTS*), bottleneck model (*agingTTS-BN*) and ground truth per one child, adult and senior speaker. Plots for all speakers can be found in Appendix 7.1

The significance of the differences between all age groups' pitch has been investigated through a one-way ANOVA statistical test in R. The ANOVA test was conducted by comparing the mean pitch of each type of synthesised sentence within the same age group. The p-values obtained were then adjusted using the Holm-Bonferroni method to counteract any effect of multiple hypotheses.

The outcome of the statistical analysis, as with every other conclusion drawn in this work, is only provisional. In the case of statistical evaluations, the size of the sample cannot be overlooked, and being my sample quite small, my findings have to be verified and studied in more depth.

The ANOVA test with the Holm-Bonferroni correction highlighted that the mean pitch of the

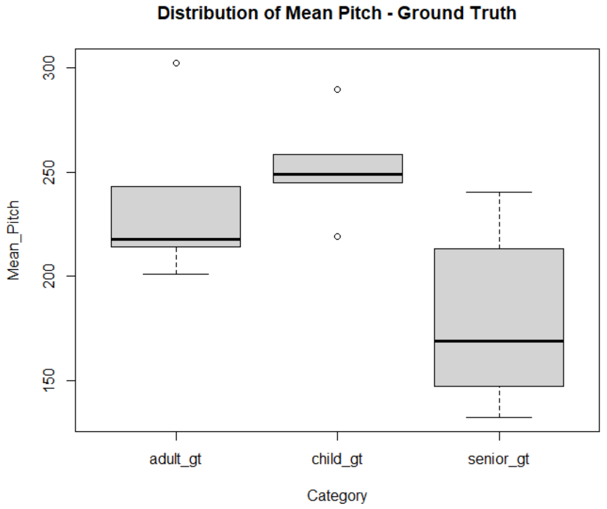


Figure 5: Mean pitch boxplot of Ground Truth audio samples

three age groups in all three scenarios is highly significant, with p-values ¡ 0.001 for the synthetic speech generated with both models (see Figure 4). On the other hand, the mean pitch is only significant with a p-value of 0.03 for the ground truth sentences (see Figure 5).

This might indicate that the two architectures accentuated differences that are not so strongly present in the ground truth in order to model the age of the speaker better. This statement has to be verified with further investigations.

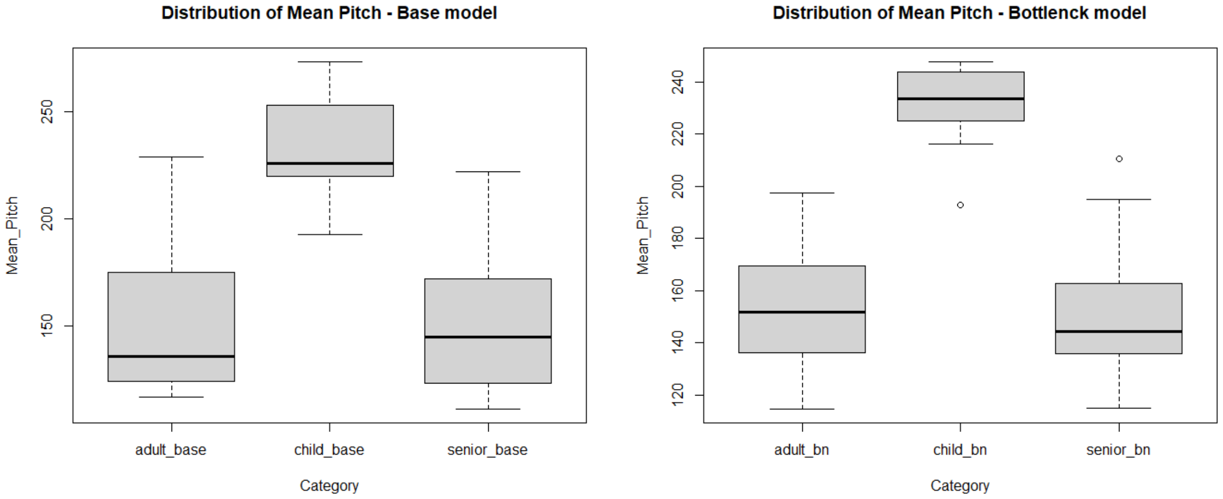


Figure 4: Mean pitch boxplot of synthetic speech from *agingTTS* and from *agingTTS-BN*

Following Barkana and Zhou (2015), I also tested the pitch range across the three age groups (see Figures 6 and 7). Even though the authors found pitch range to be a distinctive age-related feature, my analysis did not highlight statistical significance in pitch range differences in any model, as well as in the GT. In all cases, the p-value was higher than 0.1 (p-value=0.712 for *agingTTS*;

p.value=0.189 for *agingTTS-BN*; p-value=0.374 for GT). Similar to what I reported for the mean pitch, the significance of the pitch range should be further investigated with a larger sample.

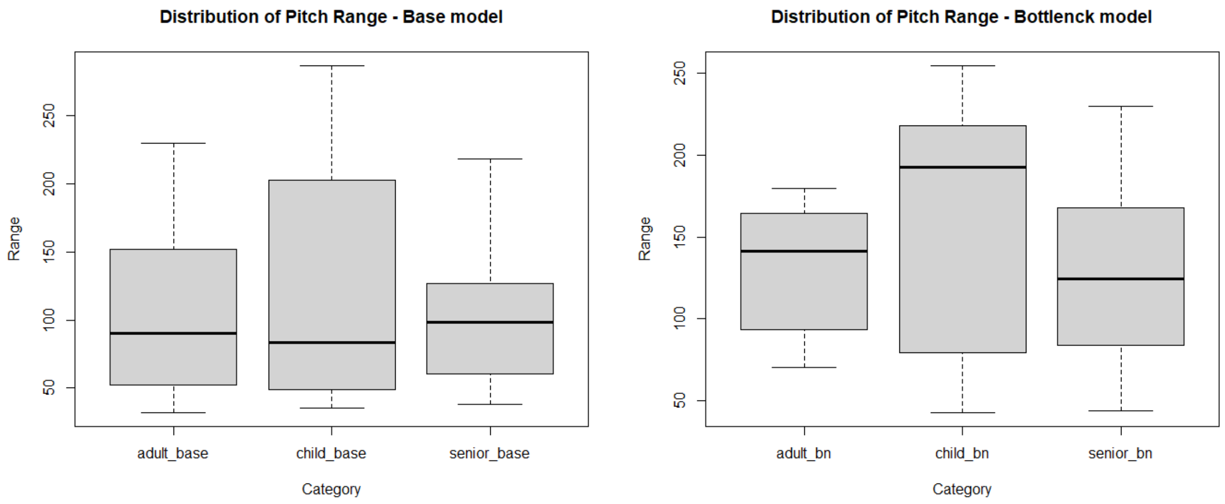


Figure 6: Pitch range boxplot of synthetic speech from *agingTTS* and from *agingTTS-BN*

4.1.2 Speaking Rate

The speaking rate was extracted using the Syllable Nuclei Praat script by Nivja H. de Jong and Heeren (2021). The script detected the nucleus of each syllable and computed the speech rate as syllable per second.

As already mentioned, given the literature on age-related voice features, the expected outcome of such analysis, for both the GT and the synthesised speech, was:

- Highest speech rate for children's speech;
- Lowest speech rate for senior's speech

The results for the selected sentences are shown in the bar plots in Figure 8, which represents the speaking rate of sentences synthesised using the base model, the model with bottleneck and finally the ground truth speaking rate.

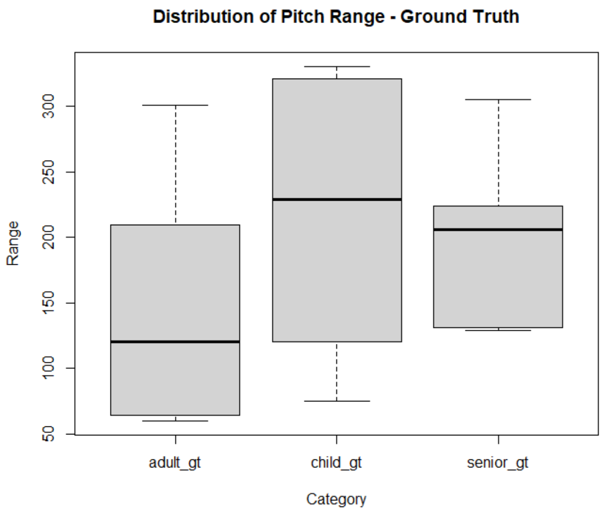
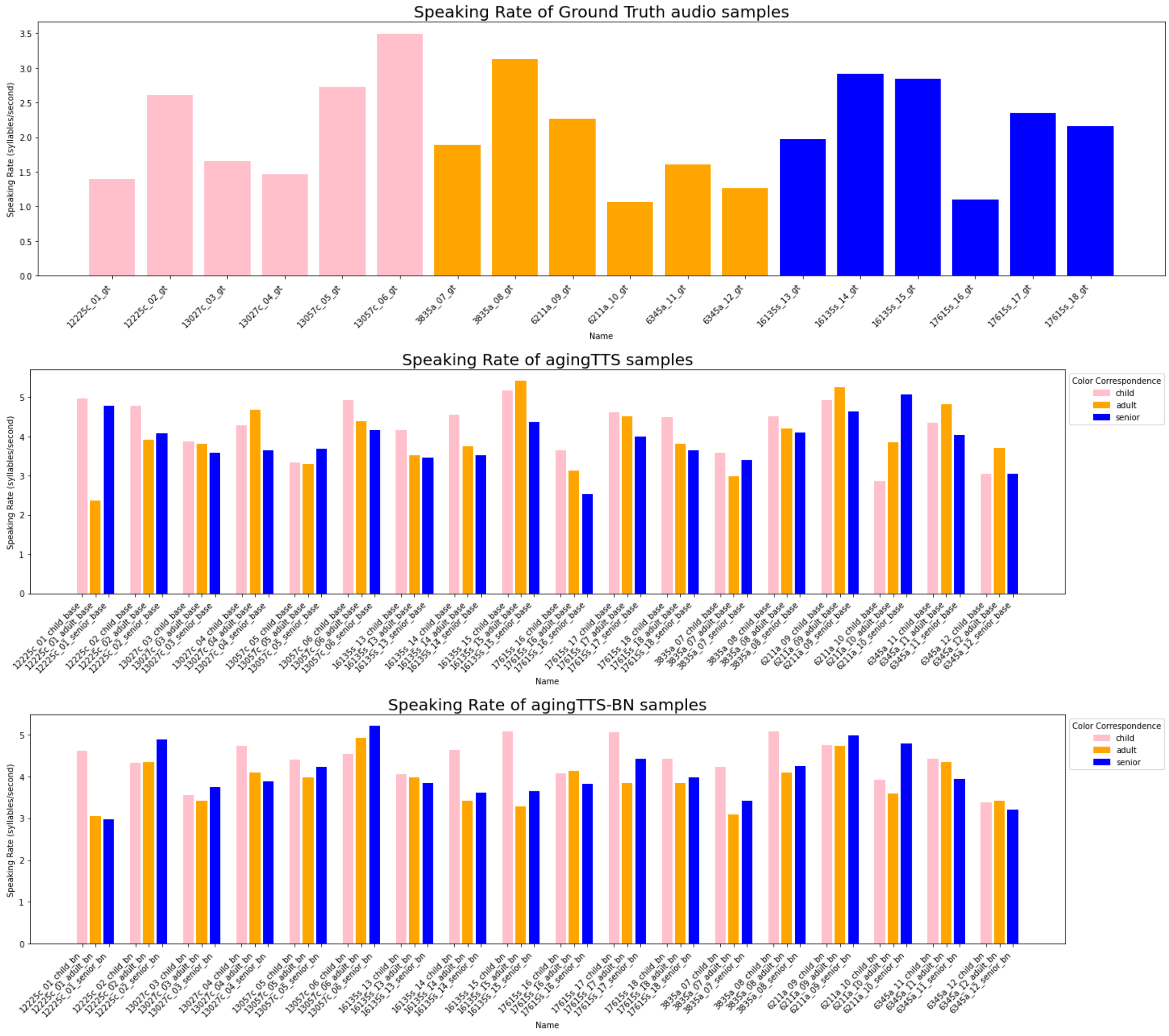


Figure 7: Pitch range boxplot of Ground Truth audio samples


Speaking Rate of Ground Truth audio samples

Speaking Rate of agingTTS samples

Speaking Rate of agingTTS-BN samples

Figure 8: Speaking rate of Ground Truth sentences, sentences synthesised with *agingTTS* and *agingTTS-BN*

In the bottleneck model (Figure 8, lower plot) there is a general increase in the speech rate of sentences synthesised with 'senior' age control, but again there is no clear pattern to be found.

Figure 10 provides a clearer picture of the speech rate distribution across conditions. From the box plot, we can observe how the ground truth consistently presents a lower speech rate compared to the synthesised counterpart. This might be due to hesitations, short pauses and other prosodic phenomena that the model was not able to reproduce.

However, the GT speech rate analysis reported in Figure 8 does not provide evidence of any pattern correlating with age.

Figure 10 provides a clearer picture of the speech rate distribution across conditions. From the box plot, we can observe how the ground truth consistently presents a lower speech rate compared to the synthesised counterpart. This might be due to hesitations, short pauses and other prosodic phe-

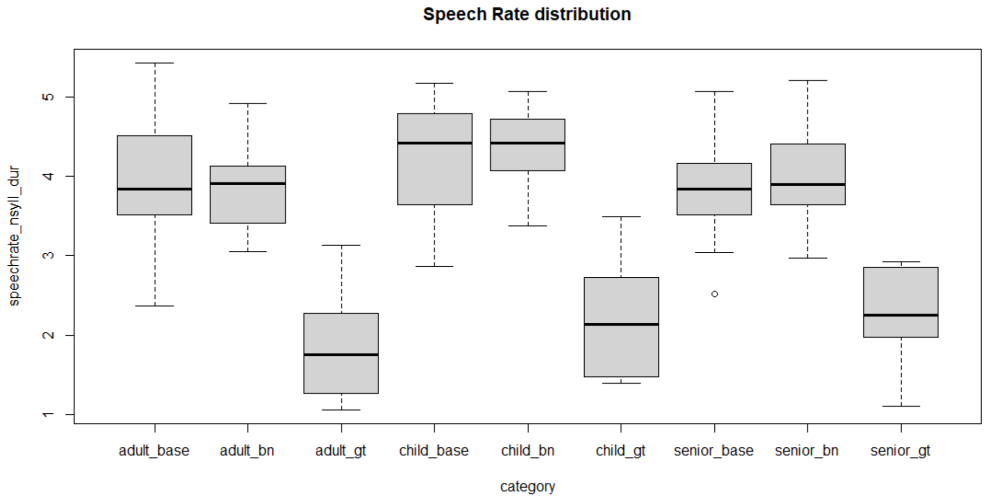


Figure 10: Boxplot for speaking rate across conditions. base = *agingTTS*; bn = *agingTTS-BN*; gt = Ground Truth

nomena that the model was not able to reproduce. An example of this can be seen in the two waveforms and spectrograms below (Figure 9), which represent sentence common_voice_en_17263012 uttered by adult speaker 6211a.

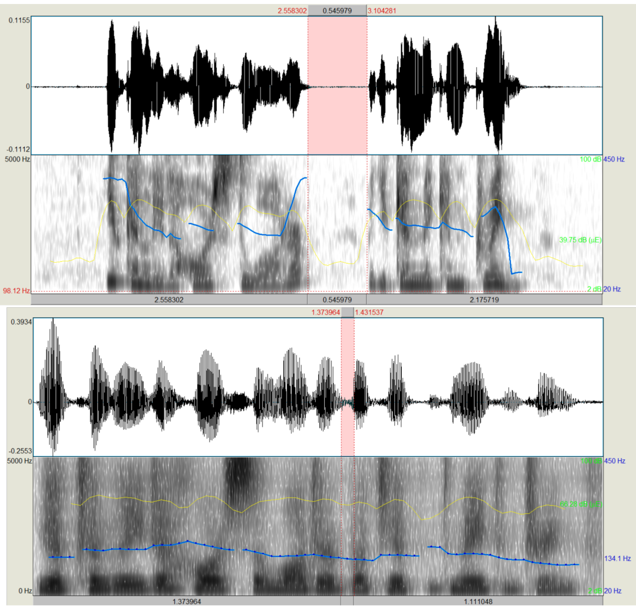


Figure 9: Waveform and spectrogram of sentence common_voice_en_17263012 uttered by adult speaker 6211a (upper image) and as synthesised with *agingTTS-BN* with age-control = 'adult' (lower image). Images extracted with Praat.

The waveform sections highlighted in light pink are pauses in the sentence, but they do not have the same length in the ground truth audio (upper plot) and the synthesised sample (lower plot). In the former, the pause is longer, around 0.5 seconds, while in the latter the same pause lasts a tenth of the original duration (0.05 second ca.).

Despite what is reported above, the GT speech rate analysis reported in Figure 8 does not provide evidence of any pattern correlating with age. The ground truth does not present any downward trend going from child to senior. This contradicts some previous findings described in Section 2.1, according to which a slower speaking rate is a primary cue in elderly speech recognition (Skoog Waller et al., 2015).

5 Ethics

The present research does not imply any specific ethical issue, since neither the research question nor the data used have self-evident and urgent ethical implications.

Even though this research does not require addressing ethical issues that are tied to its nature or data collection, some more general issues should be addressed. I believe any research that aims to improve the quality of synthesised speech to make it more similar to human production should take into consideration deepfake generation, voice cloning and their consequences. This will be addressed in Section 5.1.

Furthermore, the environmental consequences of training and using TTS and any other AI system should not be overlooked. I will consider the environmental responsibility we have as AI users and ML practitioners in Section 5.2.

5.1 Deepfakes and related issues

Even though the production of a voice closer to one's personal representation is one of the aims of this research, this may also represent a risk for this very same personal representation. In the past few years, we have had a non-negligible number of fraud cases and misuse of AI-generated voices, which on some occasions led to the ban of AI-generated voices (e.g. the banning of robocalls from the Federal Communications Commission of the USA at the beginning of 2024). One such case is the recent concerns about the use of President Biden's cloned voice for promotion purposes of the forthcoming US elections (see e.g., Seitz-Wald, 2024 and Seitz-Wald and Memoli, 2024).

Even though the TTS system outlined here is not a voice cloning system, it still goes in the direction of closing the gap between humans and AI-generated voices. As a consequence, in my opinion, together with everyone who works for this same aim, I should be mindful of this, and consider the related risks.

On the other hand, I believe the threat of misuse of AI-generated voices should not be a reason for us to stop working towards their improvement. Every tool is dangerous if used in the wrong way.

5.2 Environmental responsibility

The ethical implication of AI has recently gained attention, not only in terms of transparency and explainability of its algorithms (e.g., Floridi et al., 2018) but also in terms of its environmental impact. In the last three years, the concept of Sustainable AI (van Wynsberghe, 2021) has become more and more present in the AI discussion, since the environmental implications of training AI models cannot be ignored anymore. According to Strubell and colleagues (2019), the impact of training a deep learning NLP model has about the same carbon dioxide emissions produced by five cars over their lifetime. Since 2019, the models have grown bigger and bigger, and their impact has increased both in terms of training data and model size and the resources required to store the data and run the model have grown consequently (Wu et al., 2022).

The training of text-to-speech systems makes no exception, and working towards TTS systems with a lower ecological footprint by reducing the amount of processed data is a responsibility that every AI researcher should take.

Even though this work did not engage in the training of a TTS system with a large amount of data, it still entailed the full training of the model, which means that it still has an impact. Moreover, the system outlined here goes in the direction of working with less data, demanding less computational power to respect the environment without hindering scientific and academic advancement.

A more thorough review of methods to reduce the carbon footprint of AI and NLP systems is provided by Lacoste et al. (2019) and McDonald et al. (2022).

6 Conclusion

In this thesis, I have presented *agingTTS*, a model for age-controllable TTS, based on FastSpeech2 and integrated with an age embedding layer. The efficacy of the aforementioned model was tested by experimenting with two different setups of the age embedding layer: the baseline model, *agingTTS*, that extracts 256-dimensional age embeddings, and a model with a bottleneck in the age embedding layer, *agingTTS-BN*. Due to time constraints, no subjective evaluation was conducted on these experiments, instead, I conducted an acoustic analysis to compare the age-controlled output of the two models and the corresponding natural speech. This analysis on the one hand confirmed that the age embedding layer was able to capture some of the age-related features of the voice (e.g. differences in mean pitch), on the other hand, it is falling short of the expectations. This might be due to various limitations that the current work has. In the upcoming section (6.1), I discuss the improvement points that might lead my model to the efficacy I was aiming for.

To conclude this work, the final section will briefly reflect on the impact and potential relevance of this model in today's technological landscape.

6.1 Limitations and Future Work

As already highlighted in Chapter 4, the TTS model developed for the present thesis still has much room for improvement.

A major limitation of the present work relies upon the data used to build the training corpus. As already mentioned in 3.2, the datasets used to build the agingTTS corpus on which the model was trained are not as high-quality as a TTS system usually requires. Due to time constraints, no data enhancement method was applied to improve the data quality. This led to the output speech sounding much less natural than the usual output of FastSpeech2. Unfortunately, the limited amount of time in which this implementation was realised did not allow for a more articulated development of the system, nor for a more thorough experimentation of diverse solutions in terms of architecture.

These are the main improvement points I foresee: training data and architecture. I will discuss them in more detail throughout the current Chapter.

6.1.1 Data enhancement

A change in the training data as a whole would certainly be beneficial, since, as mentioned earlier, the datasets used were not suited for a TTS task.

The best option would be to collect data for this specific purpose, but since data collection is costly in terms of time, energy and often money, there are other solutions to the issue.

One of them is the adoption of data enhancement techniques. Particularly, I believe noise reduction techniques could successfully be applied to all the datasets employed. Being all crowdsourced and not studio-quality recordings, all the data contained background noise and other sources of disruption in the speech signal. By taking out such noisy signals, the output speech would be improved and reach a result that can be as good as the one obtained with clean data (Valentini-Botinhao and Yamagishi, 2018).

In the development of speech technologies for LRLs, to obviate the problem of the lack of data, data augmentation techniques are often employed. Given the purpose of the TTS model, it is unfortunately not possible to adopt many of these techniques. Those data augmentation methods that feature modifications in F0 and duration have in fact to be discarded. As explained in Section 2.1, pitch and speech rate are two fundamental cues for age detection in human speech.

This said, some approaches aim to increase the number of data available without changing such acoustic features. One viable option is voice cloning, which has been shown to be possible even with low-quality data (Arik et al., 2018). Nonetheless, this approach has a huge downside in the light of the present work. Voice cloning, as of now, cannot be effectively applied to a low amount of data, which means that in the present case, it might not be possible to effectively adopt this strategy. However, there has been an effort in the direction of enabling voice cloning for LRLs, such as Radhakrishnan et al., 2024 which shows promising results in this direction.

6.1.2 Improvements in the model

Testing the model outlined in this thesis with cleaner and improved data would shed light on its weaknesses at the architecture level. Nonetheless, some adjustments could be made regardless of the data used.

First of all, using different input features might be useful. As shown by Staib et al. (2020) and Do et al. (2023b), the use of input features different than phoneme labels has a non-negligible effect on the performance of a TTS model. Especially the use of articulatory features is of great impact in the case of low-resource languages, and more widely, for models working with little data (Do et al., 2023b). Since this is the case for the current model, given the fact that, as mentioned above, suitable data were scarce, the use of articulatory features instead of phoneme labels is expected to improve the performance.

Given the high individual variance and the inter-age group variance, the performance of the system might be improved by relabelling the data based on the perceived age group, and not on the actual age group. This could be achieved by training a classifier on a selected and validated subset of data. The selection of the best possible data would be done manually, while the validation could be either done manually by the developer or, for a less biased result, by a crowdsourced evaluation. The critical point of this approach might lie in the assumption that either the developer or the participants in the bottom-up evaluation are perfectly able to distinguish and classify voices based on their age. Unfortunately, this is not always the case, as discussed in 2.1, hence, the bottom-up evaluation needs to be done on a pre-selected subset of data, which raises the methodological question of how to establish who has the skills to conduct such a pre-selection. We have no other option that relies on our intuition as language users, but if we take up the suggestion made by Mei and Min, 2018 on cultural and language differences in the externalisation of age-related cues, we have to take a step further and involve native speakers in the process[13].

6.1.3 Evaluation method

The evaluation method is also quite weak and can be improved. For the subjective evaluation, the approach I foresee involves a listening test with a survey. The survey would aim to collect information about the participants in terms of their sociolinguistic background and demographic data, while the second consists of listening to audio clips of synthesised speech, 3 to 5 seconds long. Given the limitations of Mean Opinion Score (MOS) and similar evaluation methods highlighted recently

[13] This, in turn, could raise additional questions, but I won't dwell on such a complex theoretical and methodological discussion

(Wagner et al., 2019; Le Maguer et al., 2024), I would opt for a more descriptive method, following the user-centered approach proposed in Wagner et al. (2019). The overall aim of the listening test would be to evaluate whether the perceived age of the synthetic speech aligned with the input age parameter. To achieve this, accompanying each sound sample with a short description of the situation in which the listeners would have heard the voice would be instrumental to managing participants' expectations towards the use of the voice they are evaluating, and also to avoid unforeseen biases. To further avoid biases, questions asking a judgement about the speakers' age explicitly should be avoided, in favour of questions that refer to age as a way to identify the speaker.

The context provided with the audio should involve the inability to see the person speaking. This artifice should help participants relate to a situation in which the judgement would be done using only hearing cues.

The system's performance resulting from this subjective evaluation would be measured in terms of the match rate between the intended and perceived age, similarly to what has been done for the objective evaluation (see **??**).

While I hope to be able to develop and implement the above improvement points, it is also my hope that these suggestions will be taken up by future research, since I believe the potential impact of a high-performing TTS system with age-control could be important. How this system can be relevant is the focus of the following section.

6.2 Impact and relevance

Even though the outcome of my system did not satisfy my expectations and has much room for improvement, I believe it is still a step towards the creation of something that will have an impact on many people's lives.

By implementing the suggested solutions and consequently enhancing the performances, I believe a system such as the one I aimed to develop can have an impact on every kind of TTS application for which a degree of customisation is required.

First of all, a TTS with age control has the potential to enhance the usability of speech-generating technologies by allowing people to have a voice that is more personal and representative of who they are. As discussed in the introduction, our voice is part of our identity, and so is our age. Losing partially or completely the ability to speak is tough, and being able to restore one's own voice as closely as possible to its original sound, I believe, can help in the process of accepting with more ease the use of such technologies. The development of the system outlined in this work, in my opinion, can contribute to making a difference in this direction.

Additionally, the voice technology industry can also benefit from such a system. It is my belief that this represents an advancement towards the creation of user-friendly TTS systems that can be customised with minimal computational effort and no technical knowledge. This might have a fruitful application in the development of vocal personas for businesses, as well as Voice Assistants. By improving this age-controllable system, speech tech companies might be able to offer a highly flexible and easily tailored tool for third parties to create their voice bot or any other application that involves synthetic speech.

Finally, the results of this research can have an impact on academic research and the wider speech research field; I think the outcome of this work is relevant in two ways. Firstly, it highlights areas of improvement in the extraction of acoustic features related to the ageing process. Moreover, this work contributes to existing knowledge about technological solutions for under-resourced voices.

By addressing the issues raised by this work, the scientific community can advance in this field and provide better solutions to a long-standing problem such as TTS for under-resourced speech types, that can in turn be beneficial for TTS for LRLs. As stressed on multiple occasions in this thesis, the development of tools and technological resources for types of speech with low data availability and under-resourced types of voices has an increasing relevance not only within academia but also in the bigger industrial landscape.

7 Appendix

7.1 Pitch contour plots

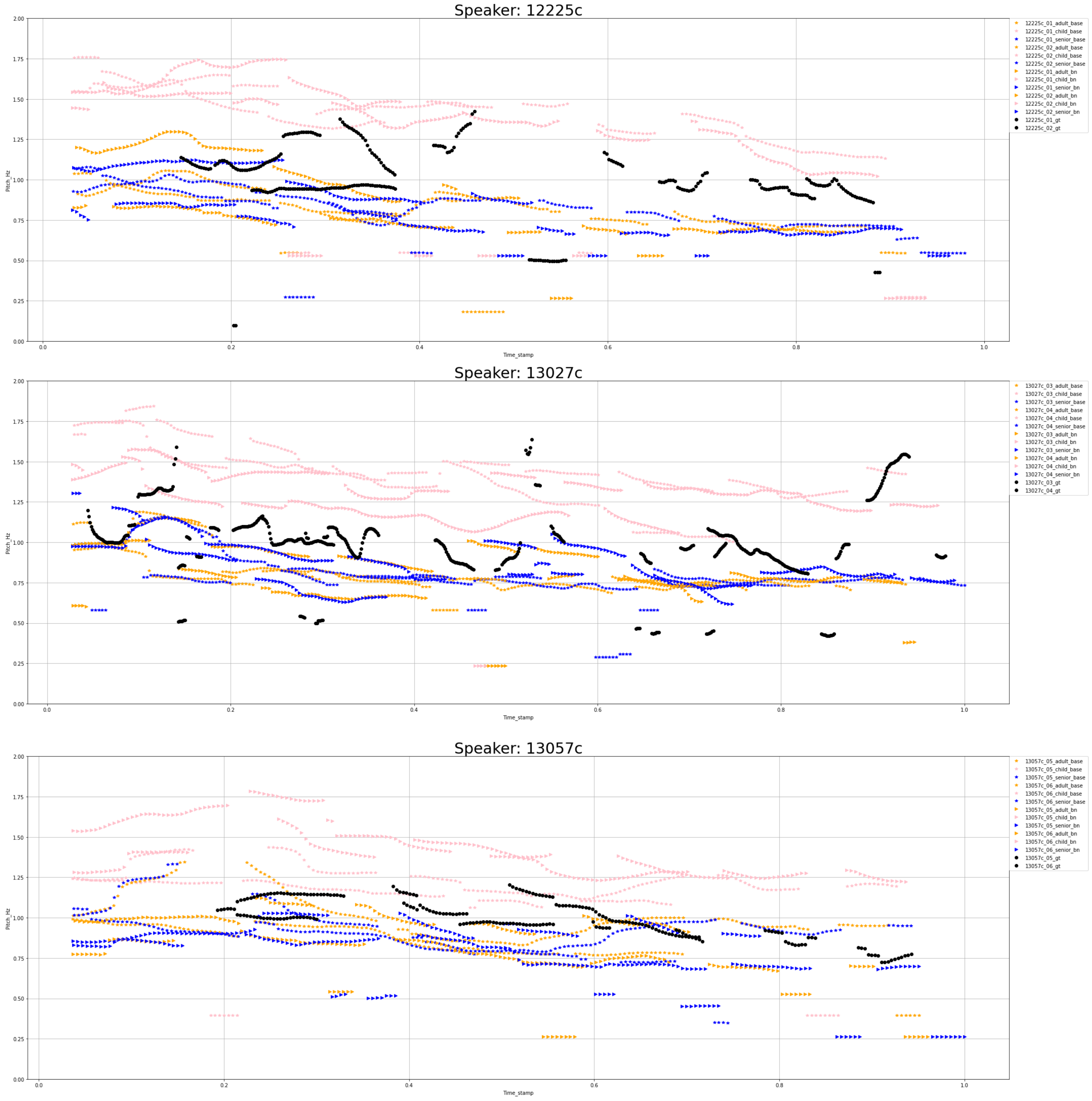


Figure 11: Aggregated pitch contour of sentences synthesised with the baseline (*agingTTS*), bottleneck model (*agingTTS-BN*) and ground truth per child speaker

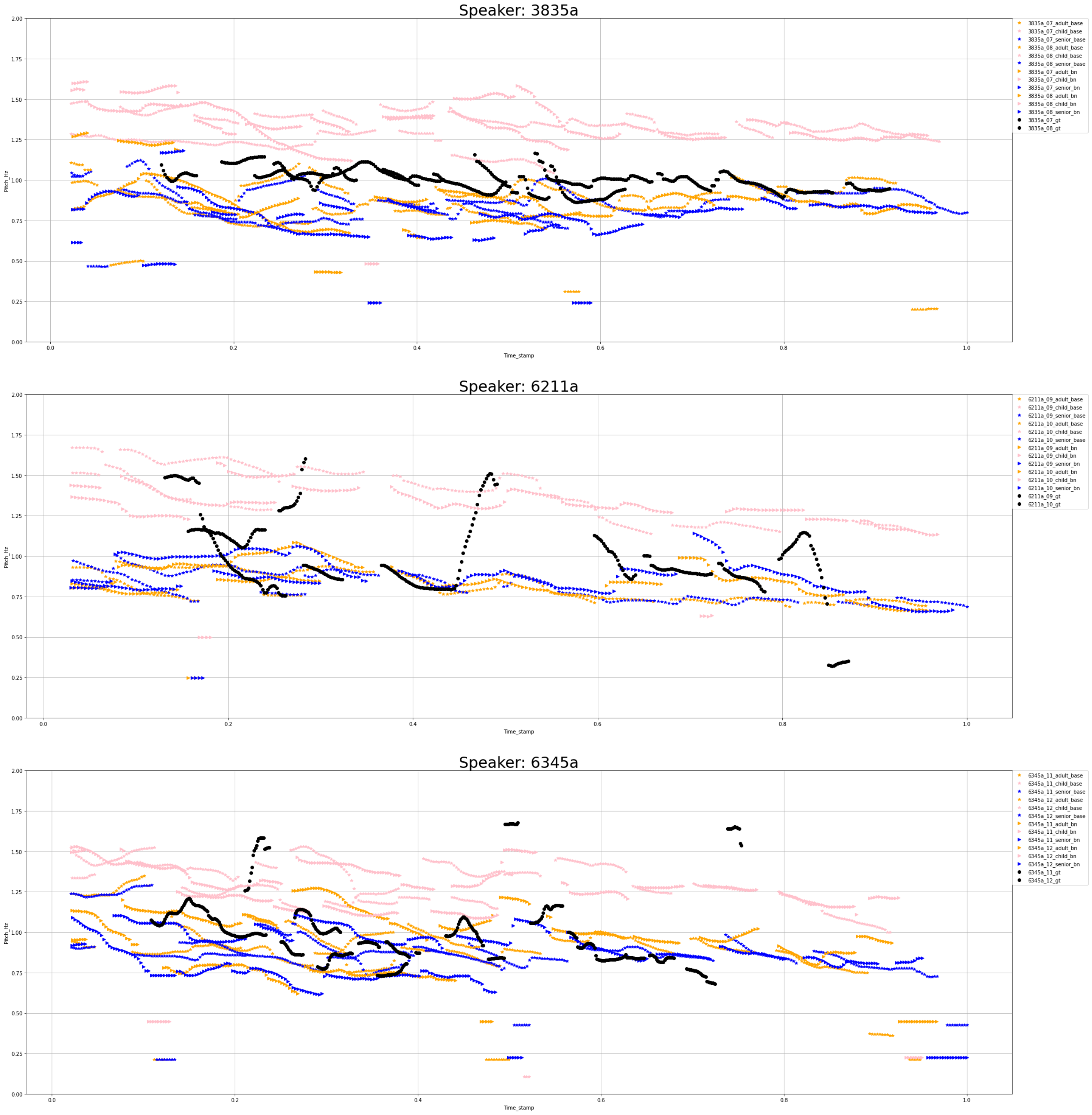


Figure 12: Aggregated pitch contour of sentences synthesised with the baseline (*agingTTS*), bottleneck model (*agingTTS-BN*) and ground truth per adult speaker

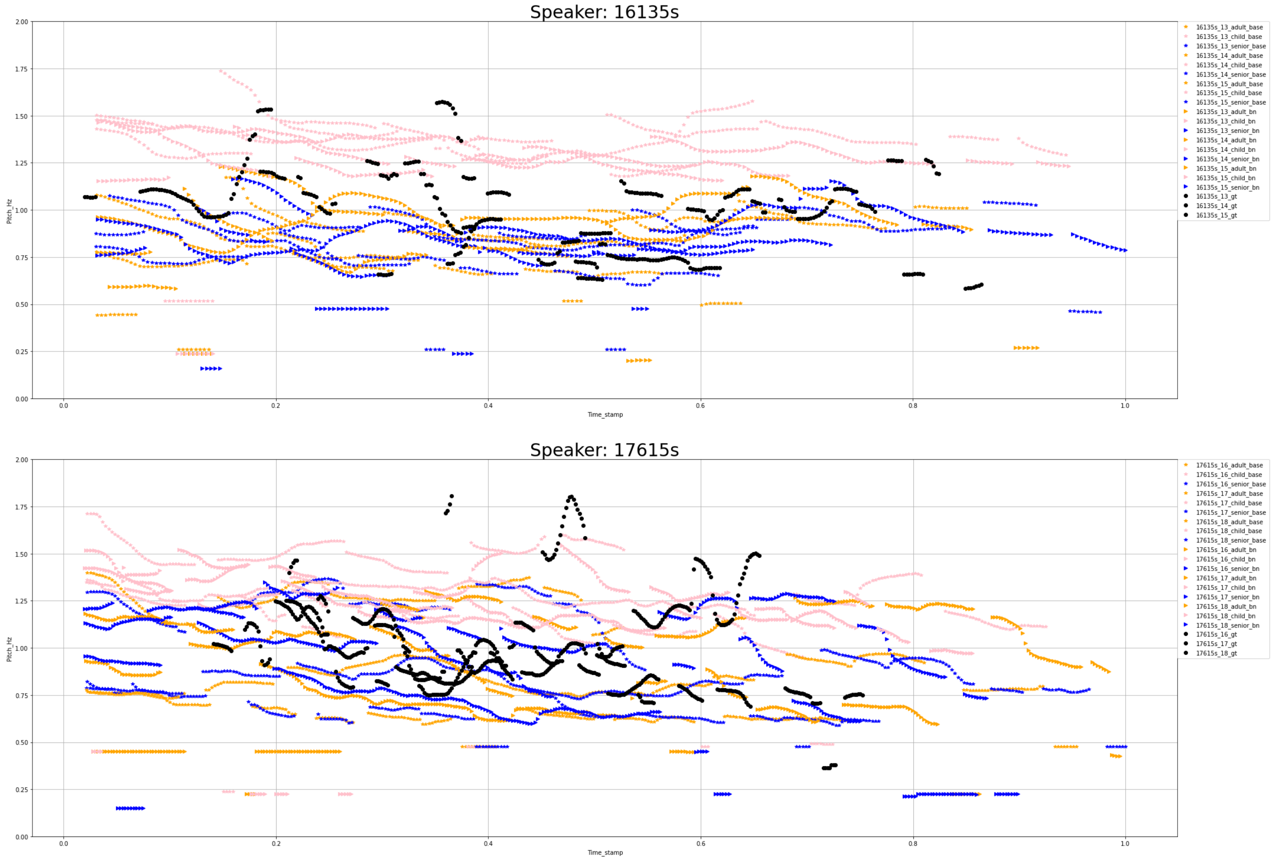


Figure 13: Aggregated pitch contour of sentences synthesised with the baseline (*agingTTS*), bottleneck model (*agingTTS-BN*) and ground truth per senior speaker

References

Ardila, R., Branson, M., Davis, K., Henretty, M., Kohler, M., Meyer, J., Morais, R., Saunders, L., Tyers, F. M., & Weber, G. (2020). Common voice: A massively-multilingual speech corpus.

Arik, S. Ö., Chen, J., Peng, K., Ping, W., & Zhou, Y. (2018). Neural voice cloning with a few samples. *CoRR*, *abs/1802.06006*. http://arxiv.org/abs/1802.06006

Barkana, B. D., & Zhou, J. (2015). A new pitch-range based feature set for a speaker's age and gender classification. *Applied Acoustics*, *98*, 52–61. https://doi.org/10.1016/j.apacoust.2015.04.013

Boersma, P., & Weenink, D. (2023). Praat: Doing Phonetics by Computer.

Burkhardt, F., Eckert, M., Johannsen, W., & Stegmann, J. (2010). A Database of Age and Gender Annotated Telephone Speech. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, & D. Tapias (Eds.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association (ELRA). Retrieved March 15, 2024, from http://www.lrec-conf.org/proceedings/lrec2010/pdf/262_Paper.pdf

Chien, C.-M., Lin, J.-H., Huang, C.-y., Hsu, P.-c., & Lee, H.-y. (2021). Investigating on incorporating pretrained and learnable speaker representations for multi-speaker multi-style text-to-speech. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8588–8592. https://doi.org/10.1109/ICASSP39728.2021.9413880

Cho, S., Nevler, N., Shellikeri, S., Parjane, N., Irwin, D. J., Ryant, N., Ash, S., Cieri, C., Liberman, M., & Grossman, M. (2021). Lexical and Acoustic Characteristics of Young and Older Healthy Adults. *Journal of Speech, Language, and Hearing Research : JSLHR*, *64*(2), 302–314. https://doi.org/10.1044/2020_JSLHR-19-00384

Davatz, G. C., Yamasaki, R., Hachiya, A., Tsuji, D. H., & Montagnoli, A. N. (2021). Source and Filter Acoustic Measures of Young, Middle-Aged and Elderly Adults for Application in Vowel Synthesis. *Journal of Voice*. https://doi.org/10.1016/j.jvoice.2021.08.025

Do, P., Coler, M., Dijkstra, J., & Klabbers, E. (2023b). The Effects of Input Type and Pronunciation Dictionary Usage in Transfer Learning for Low-Resource Text-to-Speech: Interspeech 2023. *Proceedings of Interspeech 2023*. https://doi.org/10.21437/interspeech.2023-2148

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

Eichhorn, J. T., Kent, R. D., Austin, D., & Vorperian, H. K. (2018). Effects of Aging on Vocal Fundamental Frequency and Vowel Formants in Men and Women. *Journal of Voice: Official Journal of the Voice Foundation*, *32*(5), 644.e1–644.e9. https://doi.org/10.1016/j.jvoice.2017.08.003

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, *28*(4), 689–707. https://doi.org/10.1007/s11023-018-9482-5

Gopalakrishnan, T., Imam, S. A., & Aggarwal, A. (2022). Fine Tuning and Comparing Tacotron 2, Deep Voice 3, and FastSpeech 2 TTS Models in a Low Resource Environment. *2022 IEEE International Conference on Data Science and Information System (ICDSIS)*, 1–6. https://doi.org/10.1109/ICDSIS55133.2022.9915932

Hasanabadi, M. R. (2023). An overview of text-to-speech systems and media applications. https://doi.org/10.48550/arXiv.2310.14301
Comment: Accepted in ABU Technical Review journal 2023/6.

Huckvale, M., & Webb, A. (2015). A Comparison of Human and Machine Estimation of Speaker Age. In A.-H. Dediu, C. Martín-Vide, & K. Vicsi (Eds.), *Statistical Language and Speech Processing* (pp. 111–122). Springer International Publishing. https://doi.org/10.1007/978-3-319-25789-1_11

Huff, E. W., Stigall, B., Brinkley, J., Pak, R., & Caine, K. (2020). Can Computer-Generated Speech Have an Age? *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–7. https://doi.org/10.1145/3334480.3383082

Jain, R., Yiwere, M., Bigioi, D., Corcoran, P., & Cucu, H. (2022). A Text-to-Speech Pipeline, Evaluation Methodology, and Initial Fine-Tuning Results for Child Speech Synthesis. https://doi.org/10.48550/arXiv.2203.11562

Jia, Y., Zhang, Y., Weiss, R. J., Wang, Q., Shen, J., Ren, F., Chen, Z., Nguyen, P., Pang, R., Moreno, I. L., & Wu, Y. (2019). Transfer Learning from Speaker Verification to Multispeaker Text-To-Speech Synthesis. https://doi.org/10.48550/arXiv.1806.04558

Johfre, S., & Saperstein, A. (2023). The Social Construction of Age: Concepts and Measurement. *Annual Review of Sociology*, *49*(1), 339–358. https://doi.org/10.1146/annurev-soc-031021-121020

Kuppusamy, K., & Eswaran, C. (2022). Convolutional and Deep Neural Networks based techniques for extracting the age-relevant features of the speaker. *Journal of Ambient Intelligence and Humanized Computing*, *13*(12), 5655–5667. https://doi.org/10.1007/s12652-021-03238-1

Lacoste, A., Luccioni, A., Schmidt, V., & Dandres, T. (2019). Quantifying the Carbon Emissions of Machine Learning. https://doi.org/10.48550/arXiv.1910.09700

Le Maguer, S., King, S., & Harte, N. (2024). The limits of the Mean Opinion Score for speech synthesis evaluation. *Computer Speech & Language*, *84*, 101577. https://doi.org/10.1016/j.csl.2023.101577

Lennes, M. (2003). Getting pitch maximum praat script. http://phonetics.linguistics.ucla.edu/facilities/acoustic/collect_pitch_data_from_files.txt

Linville, S. E. (1996). The sound of senescence. *Journal of Voice*, *10*(2), 190–200. https://doi.org/10.1016/S0892-1997(96)80046-4

Luong, H.-T., Takaki, S., Henter, G. E., & Yamagishi, J. (2017). Adapting and controlling DNN-based speech synthesis using input codes. *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 4905–4909. https://doi.org/10.1109/ICASSP.2017.7953089

Magueresse, A., Carles, V., & Heetderks, E. (2020). Low-resource Languages: A Review of Past Work and Future Challenges. *ArXiv*.

McAuliffe, M., Socolof, M., Mihuc, S., Wagner, M., & Sonderegger, M. (2017). Montreal Forced Aligner: Trainable Text-Speech Alignment Using Kaldi. *Proc. Interspeech 2017*, 498–502. https://doi.org/10.21437/Interspeech.2017-1386

McDonald, J., Li, B., Frey, N., Tiwari, D., Gadepally, V., & Samsi, S. (2022). Great Power, Great Responsibility: Recommendations for Reducing Energy for Training Language Models. *Findings of the Association for Computational Linguistics: NAACL 2022*, 1962–1970. https://doi.org/10.18653/v1/2022.findings-naacl.151

Mei, G., & Min, X. (2018). Automatic Age Estimation Based on Vocal Cues and Deep Neural Network. In F. Qiao, S. Patnaik, & J. Wang (Eds.), *Recent Developments in Mechatronics and Intelligent Robotics* (pp. 208–213). Springer International Publishing. https://doi.org/10.1007/978-3-319-65978-7_32

Minematsu, N., Sekiguchi, M., & Hirose, K. (2002). Automatic estimation of one's age with his/her speech based upon acoustic modeling techniques of speakers. *2002 IEEE International Conference on Acoustics, Speech, and Signal Processing*, *1*, I–137–I–140. https://doi.org/10.1109/ICASSP.2002.5743673

Moyse, E. (2014). Age Estimation from Faces and Voices: A Review. *54*(3), 255. https://doi.org/10.5334/pb.aq

Nivja H. de Jong, J. P., & Heeren, W. (2021). Praat scripts to measure speed fluency and breakdown fluency in speech automatically. *Assessment in Education: Principles, Policy & Practice*, *28*(4), 456–476. https://doi.org/10.1080/0969594X.2021.1951162

Ouchi, Y., Rakugi, H., Arai, H., Akishita, M., Ito, H., Toba, K., Kai, I., & on behalf of the Joint Committee of Japan Gerontological Society (JGLS) and Japan Geriatrics Society (JGS) on the definition and classification of the elderly. (2017). Redefining the elderly as aged 75 years and older: Proposal from the Joint Committee of Japan Gerontological Society and the Japan Geriatrics Society. *Geriatrics & Gerontology International*, *17*(7), 1045–1047. https://doi.org/10.1111/ggi.13118

Pradhan, S. S., Cole, R. A., & Ward, W. H. (2023). My Science Tutor (MyST) – A Large Corpus of Children's Conversational Speech. https://doi.org/10.48550/arXiv.2309.13347

Qawaqneh, Z., Mallouh, A. A., & Barkana, B. D. (2017). Deep neural network framework and transformed MFCCs for speaker's age and gender classification. *Knowledge-Based Systems*, *115*, 5–14. https://doi.org/10.1016/j.knosys.2016.10.008

R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Radhakrishnan, V., Aadharsh Aadhithya, A., Mohan, J., Visweswaran, M., Jyothish Lal, G., & Premjith, B. (2024). Voice cloning for low-resource languages. investigating the prospects for tamil. In *Automatic speech recognition and translation for low resource languages* (pp. 243–257). John Wiley Sons, Ltd. https://doi.org/https://doi.org/10.1002/9781394214624.ch12

Ren, Y., Hu, C., Tan, X., Qin, T., Zhao, S., Zhao, Z., & Liu, T.-Y. (2022). FastSpeech 2: Fast and High-Quality End-to-End Text to Speech. https://doi.org/10.48550/arXiv.2006.04558

Seitz-Wald, A. (2024). Democratic operative admits to commissioning fake Biden robocall that used AI. *NBC News*. Retrieved March 11, 2024, from https://www.nbcnews.com/politics/2024-election/democratic-operative-admits-commissioning-fake-biden-robocall-used-ai-rcna140402

Seitz-Wald, A., & Memoli, M. (2024). Fake Joe Biden robocall tells New Hampshire Democrats not to vote Tuesday. *NBC News*. Retrieved March 11, 2024, from https://www.nbcnews.com/politics/2024-election/fake-joe-biden-robocall-tells-new-hampshire-democrats-not-vote-tuesday-rcna134984

Skoog Waller, S., & Eriksson, M. (2016). Vocal Age Disguise: The Role of Fundamental Frequency and Speech Rate and Its Perceived Effects. *Frontiers in Psychology*, *7*.

Skoog Waller, S., Eriksson, M., & Sörqvist, P. (2015). Can you hear my age? Influences of speech rate and speech spontaneity on estimation of speaker age. *Frontiers in Psychology*, *6*, 978. https://doi.org/10.3389/fpsyg.2015.00978

Staib, M., Teh, T. H., Torresquintero, A., Mohan, D. S. R., Foglianti, L., Lenain, R., & Gao, J. (2020). Phonological Features for 0-shot Multilingual Speech Synthesis. *Interspeech 2020*, 2942–2946. https://doi.org/10.21437/Interspeech.2020-1821

Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and Policy Considerations for Deep Learning in NLP. https://doi.org/10.48550/arXiv.1906.02243

Valentini-Botinhao, C., & Yamagishi, J. (2018). Speech enhancement of noisy and reverberant speech for text-to-speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, *26*(8), 1420–1433. https://doi.org/10.1109/TASLP.2018.2828980

van Wynsberghe, A. (2021). Sustainable AI: AI for sustainability and the sustainability of AI. *AI and Ethics*, *1*(3), 213–218. https://doi.org/10.1007/s43681-021-00043-6

Wagner, P., Beskow, J., Betz, S., Edlund, J., Gustafson, J., Eje Henter, G., Le Maguer, S., Malisz, Z., Székely, É., Tånnander, C., & Voße, J. (2019). Speech Synthesis Evaluation — State-of-the-Art Assessment and Suggestion for a Novel Research Program. *Proc. 10th ISCA Workshop on Speech Synthesis (SSW 10)*, 105–110. https://doi.org/10.21437/SSW.2019-19

Wan, L., Wang, Q., Papir, A., & Moreno, I. L. (2020). Generalized End-to-End Loss for Speaker Verification.

Ward, W., Cole, R., Bolaños, D., Buchenroth-Martin, C., Svirsky, E., Vuuren, S. V., Weston, T., Zheng, J., & Becker, L. (2011). My science tutor: A conversational multimedia virtual tutor for elementary school science. *ACM Transactions on Speech and Language Processing*, *7*(4), 18:1–18:29. https://doi.org/10.1145/1998384.1998392

Ward, W., Pradhan, S., & Cole, R. (2021). MyST Children's Conversational Speech. https://doi.org/10.35111/CYXY-P432

Wells, D., & Richmond, K. (2021). Cross-lingual Transfer of Phonological Features for Low-resource Speech Synthesis. *11th ISCA Speech Synthesis Workshop (SSW 11)*, 160–165. https://doi.org/10.21437/SSW.2021-28

WHO, W. H. O. (2019). Adolescent health. https://www.who.int/health-topics/adolescent-health/#tab=tab_1

WHO, W. H. O. (2022). Ageing and health. https://www.who.int/news-room/fact-sheets/detail/ageing-and-health

Wu, C.-J., Raghavendra, R., Gupta, U., Acun, B., Ardalani, N., Maeng, K., Chang, G., Aga, F., Huang, J., Bai, C., Gschwind, M., Gupta, A., Ott, M., Melnikov, A., Candido, S., Brooks, D., Chauhan, G., Lee, B., Lee, H.-H., … Hazelwood, K. (2022). Sustainable AI: Environmental Implications, Challenges and Opportunities. *Proceedings of Machine Learning and Systems*, *4*, 795–813.

Yücesoy, E. (2023). Speaker age and gender recognition using 1D and 2D convolutional neural networks. *Neural Computing and Applications*, *36*(6), 3065–3075. https://doi.org/10.1007/s00521-023-09153-0