Final Year Project Proposal

TU858

Aria – an automated voice trainer for singers

Elitsa Koleva
C21320836

School of Computer Science

TU Dublin – City Campus

13/10/2024

Table of Contents

Declaration

I hereby declare that the work described in this dissertation is, except where otherwise stated, entirely my own work and has not been submitted as an exercise for a degree at this or any other university.

Signed:

Elitsa Koleva

Elitsa Koleva

13/10/2024

Summary

Aria is a full-stack web application designed to improve its user's singing abilities without having to get help from professional voice coaches.

The app will analyse the user's voice through a vocal analysis test, gathering information on pitch range, resonance, formant frequencies, tone quality, breath control, vocal stability, emotional expressiveness and vocal endurance.
Using this data, Aria will generate a personalised exercise plan tailored to improve the user's unique voice, with the ability to practice daily and track progress through weekly vocal assessments.

Additionally, Aria will offer theoretical lessons that complement the exercises, enhancing the user's understanding of vocal techniques and vocal health.

This project offers a convenient and affordable, personalised solution for vocal training and aims to deliver a comprehensive tool that users can use to improve their voices at their own pace.

Background (and References)

Voice Production

The human voice relies on respiration, phonation, and resonance to produce sound.

Respiration provides the necessary air pressure for voice production, proper control of breath is critical in singing, as it figures out the stability and volume of the voice. Without sufficient air pressure, phonation (sound production) cannot occur effectively [5].

Phonation is the process of producing sound by the vibration of the vocal folds (vocal cords). The vocal folds, found in the larynx, chop the steady airflow into a series of quasi-periodic air pulses. This produces the fundamental tone of the voice, characterised by a spectrum of harmonic partials, which are multiples of the fundamental frequency. As frequency rises, the amplitudes of these partials decrease monotonically, creating a rich, harmonic structure [5].

Resonance shapes the sound in the vocal tract. When the airflow pulses from the vocal folds pass through this tract, they are changed by its shape and size. The vocal tract has resonances known as formants, which amplify certain frequencies [5].

Formants and Harmonics

Formants are peaks in the vocal tract's resonances that enhance nearby partials and shape vowel quality. The two lowest formants, F1 and F2, are critical for vowel differentiation and can be changed over a range of two octaves or more while higher formants cannot be varied as much and do not contribute to vowel quality, they signify personal voice timbre [5]

Pitch (F0) is decided by the fundamental frequency of the voice, which stays stable even if higher harmonics are altered [5].

H1-H2 is a parameter that measures the difference in amplitude between the first and second harmonics. It reflects the degree of glottal adduction, influencing voice quality from breathy to pressed tones [5].

Vocal Registers like vocal fry, modal, and falsetto in men, or chest, middle, head, and whistle in women, reflect different modes of vocal fold vibration and cover various parts of the pitch range [5]

Transitioning between registers is often accompanied by a change in pitch, and trained singers aim to minimise the timbral differences between registers [5].

Health and Singing

Singing offers various health benefits, such as improving mental well-being and vocal health, as well as enhancing social interaction [8].

Despite all the benefits many people around the world, especially in rural areas, have no practical access to high quality voice instruction due to unavailability of qualified coaches or the associated cost with getting tutoring. [1]

Improper training can lead to vocal damage, highlighting the need for accessible vocal coaching.

Vocal Pedagogy and Existing Technology

Vocal pedagogy has a wide range of conflicting views, particularly when distinguishing between classical and pop singing [2].

Pop singing is a genre of singing with varying styles [7].
There is extraordinarily little research that defines what the technicalities of pop singing are, as it is learned often informally and trough imitation [3].
There are some studies showing that, compared to classical singing or blues singing, pop singing' phonation is neutral (i.e. close to natural speech) [6].

When it comes to existing solutions, several apps are available for vocal coaching, such as Vocal Image and Yousician.

Vocal Image - AI vocal coach

This app uses AI to help people improve their voice in a wide range of areas including public speaking, voice feminisation/ masculinisation for people in the LGBTQ+ community as well as speech recovery and singing training.
The app offers a wide range of tests[1] that the user can take and see relevant results.
For example, the Voice Test asks the user to read out a short paragraph, after which it gives a percentage rating of the user's voice's masculinity vs femininity, confidence versus weakness

[1] "Voice Rating – find out how your voice sounds like to others, Archetype Test - discover which archetype your voice belongs to, Voice Test – analyse your voice pitch & volume in numbers, Accent Test – get insights about your accent, Age Test – find out how old your voice sounds like, Clarity Test – check your diction accuracy, Singing Test – discover your voice type, pitch and volume range and Celebrity Test – compare your voice to the voices of TOP stars." (Vocal Image App – Home page, 2024)

and how monotone the voice sounds. It also offers pitch related insights (i.e. average pitch, median pitch, and highest/lowest pitch) and volume information (i.e. average volume, median volume, and highest/lowest volume in decibels).
Their Singing Test asks the user to sing anything for 10-15 seconds, after processing, the results provide information about the user's vocal range in octaves, volume range, voice type (i.e. soprano, mezzo-soprano, baritone, alto, tenor, and bass), and an associated pitch graph.

Upon personal investigation, Vocal Image seems to focus on public speaking and confidence rather than singing. The testing parameters are quite limited and would not be sufficient for a full in-depth vocal analysis. The UI/UX also has space for improvement as the user must choose tests by themself and choose their own exercises.
The app also provides an AI chat bot that is meant to give recommendations to the user on what exercises to do based on their data.

Vocal Image provides educational content as well in the form of podcasts, videos, literature, and short form content to help users improve their voice.

When it comes to the pricing of the app, it is using a subscription model with normal pricing.

Yousician
Yousician offers a wide range of plans for learning different musical instruments, one of which is learning how to sing.
The singing plan uses imitation type learning [4] where the user tries to sing along a song and sees live feedback of their pitch combined with on-screen lyrics.
The app does not provide detailed voice analysis or personalised detailed training the way a traditional voice coach would, and it is also a subscription-based service.

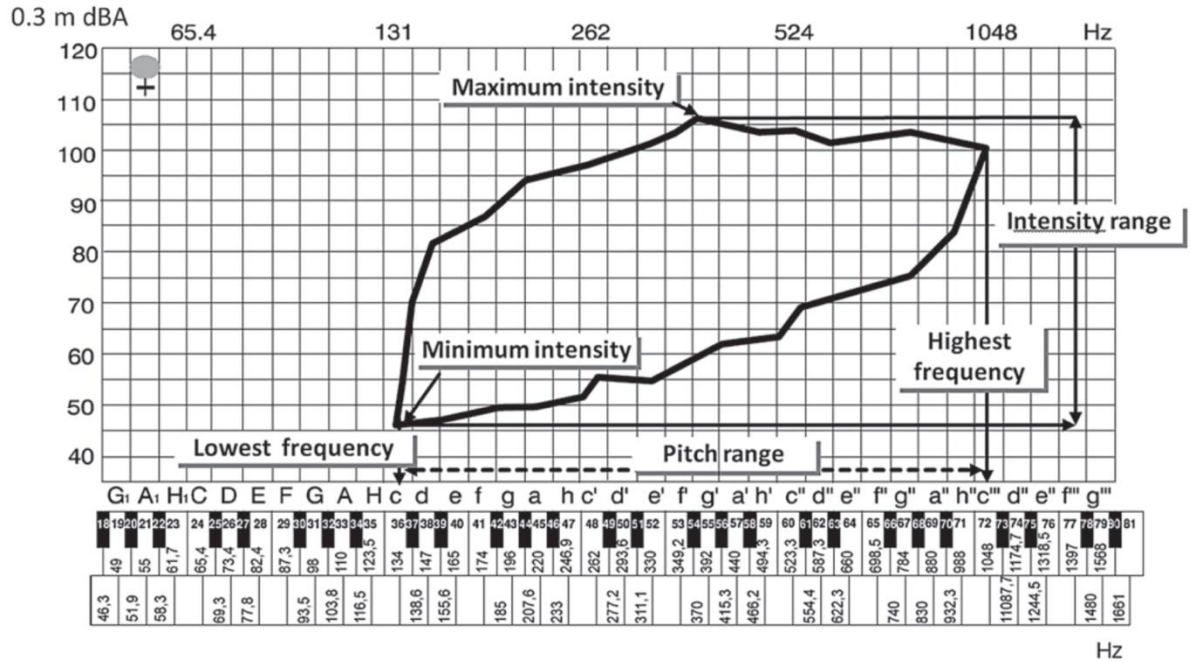
To create Aria's stand-out feature, the in-depth voice analysis, I need to be able to gather information on the respiration, phonation, and resonance of the user's voice.

Based on the existing competition, an app like Aria that provides in-depth vocal analysis and personalised exercises for singing is missing from the market.

Voice Analysis Technologies

Pitch-Range and Pitch Accuracy

Voice Range Profiles (VRP) are commonly used to illustrate the dynamic range of a singer's voice by measuring the lowest and highest frequencies they can produce and their intensity. In a study examining the VRP of singing students, it was found that healthy female singers typically achieved a low frequency of around 134 Hz (C3) and a high frequency of approximately 1048 Hz (C6) [15].



While trained singers often have a vocal range spanning 3 to 4 octaves, the average person's vocal range is typically narrower, usually covering around 1.5 to 2 octaves. To assess someone's vocal range, we can use their natural speaking pitch as a midpoint and measure their ability to reach two octaves above and below this median. This method provides a clear evaluation of the person's vocal-range potential.

In the field of voice analysis, several algorithms exist for detecting pitch, such as Fast Fourier Transform (FFT), Harmonic Product Spectrum (HPS), and the YIN algorithm. These methods allow for precise analysis of pitch variations in audio samples. Libraries like Librosa offer comprehensive tools for audio analysis, including pitch detection and voice range profiling, making them valuable resources for the purposes of this project.

Vocal Timbre

Timbre is the soul of the voice (i.e. the identifying characteristics of one's voice, for example, when the same note is played on two different instruments e.g. Piano and Trumpet, we can tell which one is which).
It cannot be changed as it is related to the physical structure of a person's body [13].

As timbre is based on how the listener perceives the voice, it can be hard to define a set of descriptors for it. Madeleine Harbey suggests defining vocal timbre by listening to a sample of natural speech and describing it in the following set of metaphors, dark or bright, dull (breathiness and smokiness) or clear, heavy or light, fluty or brassy (nasally) [13].

A study looking at different ways to describe vocal timber found that vocal timbre can be described through various parameters that capture both the physiological production of sound and its perceptual qualities. [14]
Key descriptors include the phonation onset, which can range from "*breathy"* or "*creaky" to* "*smooth"* or "*glottal"*, and the position and shape of the vocal tract, influencing resonance and formant frequencies. [14]
Adjustments in the vocal tract such as the constriction or expansion of the pharynx, larynx positioning, and velum movement, lead to distinct timbres like "oral twang", "nasal twang", or "sob," each characterised by specific acoustic and physiological traits. [14]

Lastly, vocal timbre can be modulated by breath support and muscular anchoring, where high airflow and muscular engagement often accompany powerful timbres like belting, while softer sounds like falsetto and breathy voices involve lower pressure and airflow. [14]

A machine learning model can be trained on a set of labelled data that includes a natural speech sample, descriptors, and higher formant frequency ranges to be able to predict the user's vocal timbre.
The frequency data can be captured using FFT and added as a descriptive feature to the predictive model.

Breath Control

Breath control is a vital aspect of singing, especially in pop music, where maintaining consistent vocal strength and tone across various melodies is essential. It involves managing the airflow through the vocal cords in coordination with the diaphragm to regulate pitch, volume, and the ability to sustain long notes or sing intricate phrases.

Improper breath control can cause shortness of breath, gasping for air during inhalation, forced expiration, reduced airflow during phonation, and beginning and terminating phonation at lower-than-normal lung capacity [16].

To be able to analyse breath control and track its progress two measures will be used.
Lung capacity will be evaluated by recording the time taken for the user to fully inhale and exhale.
Note sustain duration will also be recorded by asking the user to sing a note for as long as they can.

Using these two simple measures, breath control progress can be tracked and analysed over time.

Vocal Stability

To be able to detect vocal stability, the consistency of frequency and amplitude can be assessed using existing algorithms for calculating absolute and relative jitter and shimmer:

$$Jitter(absolute) = \frac{1}{N-1}\sum_{i=1}^{N-1} \| T_i - T_{i+1} \|$$

$$Jitter(relative) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1} \| T_i - T_{i+1} \|}{\frac{1}{N}\sum_{i=1}^{N} T_i}$$

$$Shimmer(dB) = \frac{1}{N-1}\sum_{i=1}^{N-1} \| 20\log\left(\frac{A_{i+1}}{A_i}\right) \|$$

$$Shimmer(relative) = \frac{\frac{1}{N-1}\sum_{i=1}^{N-1} \| A_i - A_{i+1} \|}{\frac{1}{N}\sum_{i=1}^{N} A_i}$$

Jitter and Shimmer are most clearly detected for long, sustained vowels [11].

Emotional Expressiveness

Sentiment Analysis using ML models can be used to classify emotional tones using labelled data (e.g. neutral, happy, sad, angry).

Vocal Endurance

This can be calculated by analysing the vocal fatigue and consistency over extended periods of use.

Singing voice datasets

VocalSet is a singing voice dataset consisting of 10.1 hours of monophonic recorded audio of professional singers demonstrating both standard and extended vocal techniques on all 5 vowels [17].

Proposed Approach

Methodology

I will be using the Scrum methodology to complete Aria.

Scrum is an agile project management framework that provides structure for teams to manage their work through a set of values, principles and practices [9].

It works by organising the tasks needed to complete a project into a scrum artefact called a Product Backlog.
Then some of the tasks from the Product Backlog are selected to be developed during a period (i.e. two weeks), called a Sprint.
After a Sprint is completed, the results are tested, and the next Sprint gets planned.
This process is repeated until the full Product Backlog is completed [10].

Scrum is based on empiricism[2] and lean thinking and employs an iterative and incremental approach to complete work [10].

Since this project is individual work, some of the Scrum figures and practices will be excluded. For example, the Scrum Master, Scrum Product Owner and Scrum development team will all be the author.
Sprint reviews (i.e. a meeting at the end of a sprint to show and discuss the work done during the sprint [9]) and Sprint planning (i.e. a meeting held at the beginning of a sprint to identify which tasks should be completed during the sprint [9]) will be conducted in collaboration with the supervisor.
Daily stand-ups (i.e. a short daily meeting of the scrum development team, scrum master and product owner to discuss progress [9]) will be omitted.

[2] Fumerton, R., Quinton,. Anthony M., Quinton, Baron and Duignan, Brian, 2024, empiricism, Encyclopaedia Britannica. https://www.britannica.com/topic/empiricism

Design and Research

User Requirements Gathering

Target Audience

The target audience for a traditional voice coach would primarily be amateur and professional singers, voice actors, people looking to improve their vocal health (e.g. speech therapists, public speakers) and individuals aiming for voice feminisation/masculinisation.

Since Aria's exercise plans are going to be focusing on singing training, the target audience would primarily include amateur and professional singers.

The target audience could potentially be expanded by utilising the voice analysis results. Voice actors, people looking to improve their vocal health (e.g. speech therapists, public speakers) and individuals aiming for voice feminisation/masculinisation can use the voice analysis test to track their progress. The theoretical content would also be useful for all groups.

Product Backlog

ID	Task Name	PRIORITY	SPRINT	STATUS
101	Vocal Analysis Test UI/UX	Medium		Not Started
1021	Vocal Analysis Test - Posture and Physical Alignment Check	Medium		Not Started
1022	Vocal Analysis Test - Pitch Range Test	High		Not Started
1023	Vocal Analysis Test - Pitch Accuracy Test	High		Not Started
1024	Vocal Analysis Test - Breath Control/ Lung Capacity Test	High		Not Started
1025	Vocal Analysis Test - Tone and Timbre Test	High		Not Started
1026	Vocal Analysis Test - Vocal Agility Test	Low		Not Started
1027	Vocal Analysis Test - Emotional Expressiveness	Low		Not Started
1028	Vocal Analysis Test - Vocal Endurance	Medium		Not Started
103	Vocal Analysis Test Results AI Interpretation	High		Not Started
104	Vocal Analysis Test Results Exercise recommendations	High		Not Started
105	Vocal Analysis Test Results Download	Low		Not Started
201	Vocal Exercises for Pitch Accuracy	Medium		Not Started
202	Vocal Exercises for Strengthening Vocal Resonance	Medium		Not Started
203	Vocal Exercises for Improving Tone Quality	Medium		Not Started
204	Vocal Exercises for Emotional Expressivness	Low		Not Started
205	Vocal Exercises for Breath Control	Medium		Not Started
206	Vocal Exercises for Vocal Range and Flexibility	Medium		Not Started
207	Vocal Exercises Warm Ups	Medium		Not Started
301	Theory Lessons - Breathing Techniques	Medium		Not Started
302	Theory Lessons - Vocal Anatomy	Medium		Not Started
303	Theory Lessons - Pitch and Intonation	Low		Not Started
304	Theory Lessons - Vocal Registers	Low		Not Started
305	Theory Lessons - Tone and Timbre	Low		Not Started
306	Theory Lessons - Vocal Health	Medium		Not Started
307	Theory Lessons - Performance Technique	Medium		Not Started
308	Theory Lessons - Song Interpretation	Medium		Not Started
309	Theory Lessons - Dynamics and Expression	Medium		Not Started

Implementation

Implementation of the app will be done in cycles of sprints.

Initial sprints will focus on creating the base frontend design for the app that will be based on the Figma design.
After the base frontend has been made, the core functionality of the app will be developed in order of priority, for example, the voice analysis test, then the exercises, then the algorithm for recommending exercises etc.

Sprint cycles will continue until the full project backlog has been completed.

Testing and Maintenance

Testing will be ongoing throughout the development process, starting with unit testing for individual components at the end of each sprint and progressing to integration testing.

User testing will also be conducted for the interim prototype and before final submission of the app.

Deliverables

The project end products include:

- Interim report
- a fully working prototype of Aria that will be able to analyse the user's voice and display the results with some text-based recommendations on how to improve the user's singing
- Final FYP report
- A fully working project, that would be able to analyse the user's voice, display their results and offer an on-site plan with exercises they can do daily to improve. The app would also have theoretical lessons and be able to track the user's progress.

Technical Requirements

For the frontend, I will be using Next.js to handle server-side rendering and API routes, Tailwind CSS for styling and Web Audio API for capturing and analysing audio in real time.

For the backend I will be using Firebase for the user authentication and real-time database, Node.js, and Python for the machine learning and AI features.

Other tools I will be using include Figma for the UI/UX design and Azure DevOps for project management.

Conclusion

Aria is an innovative approach to vocal training that provides personalised, accessible and affordable singing training to anyone who has access to the internet through a combination of real-time voice analysis, tailored exercises, and educational content.\

It will be completed using the Scrum methodology and technologies such as Next.js, Firebase and Python.

References

[1] V. Vinze, J. Dhami, D. Desai, H. Dalvi, and P. Raut, "Application Of AI As Singing Trainer," IEEE Xplore, 2021. [Online]. Available: https://ieeexplore.ieee.org/abstract/document/9697188. [Accessed: 11-Oct-2024].

[2] E. T. Craigo, Neural Network-Aided Audio Processing for Automated Vocal Coaching, Bachelor's thesis, Harvard College, 2019. [Online]. Available:

https://dash.harvard.edu/bitstream/handle/1/37364655/CRAIGO-SENIORTHESIS-2019.pdf?sequence=1&isAllowed=y. [Accessed: 12-Oct-2024].

[3] S. S. Santos, et al., "Singing style, vocal habits, and general health of professional singers," International Archives of Otorhinolaryngology, 2019. [Online]. Available: https://www.scielo.br/j/iao/a/gnhD4YMc6qBpbYWQqH5D5yp/?lang=en. [Accessed: 02-Oct-2024].

[4] G. F. Welch, D. M. Howard, and J. Nix, Eds., The Oxford Handbook of Singing, Illustrated ed., Oxford: Oxford University Press, 2019.

[5] D. Deutsch, The Psychology of Music, 3rd ed., Amsterdam: Elsevier, Academic Press, 2013.

[6] M. Thalen and J. Sundberg, "Describing different styles of singing: A comparison of a female singer's voice source in 'Classical', 'Pop', 'Jazz' and 'Blues'," Logopaedics Phoniatrics Vocology, vol. 26, no. 2, pp. 82-93, 2001, doi: 10.1080/140154301753207458.

[7] W. Li, "Analysis of the use of pop singing in musical theater singing based on data analysis," Applied Mathematics and Nonlinear Sciences, vol. 9, no. 1, 2023. doi: https://doi.org/10.2478/amns.2023.2.00867.

[8] G. Kreutz, S. Bongard, S. Rohrmann, D. Grebe, H. G. Bastian, and V. Hodapp, "Does singing provide health benefits," in Proc. 5th Triennial ESCOM Conf., Hanover, Germany: Hanover University of Music and Drama, 2003, pp. 8-13.

[9] C. Drumond, "What is scrum and how to get started," Atlassian, 2024. https://www.atlassian.com/agile/scrum

[10] K. Schwaber and J. Sutherland, "Scrum Guide," Scrumguides.org, Nov. 2020. https://scrumguides.org/scrum-guide.html

[11] "3.14. Jitter and shimmer — Introduction to Speech Processing," speechprocessingbook.aalto.fi. https://speechprocessingbook.aalto.fi/Representations/Jitter_and_shimmer.html

[12] Dat Tran-Anh, Nam Hoang Vu, Khanh Nguyen-Trong, and C. Pham, "Multi-task learning neural networks for breath sound detection and classification in pervasive healthcare," vol. 86, pp. 101685–101685, Oct. 2022, doi: https://doi.org/10.1016/j.pmcj.2022.101685.

[13] Madeleine Harvey, "Test Your Vocal Timbre," YouTube, Jun. 18, 2024. https://www.youtube.com/watch?v=4RYvTdxtC3U (accessed Oct. 20, 2024).

[14] Heidemann, Kate. "A System for Describing Vocal Timbre in Popular Song." Music Theory Online 22.1 (2016). https://mtosmt.org/issues/mto.16.22.1/mto.16.22.1.heidemann.pdf

[15] Hugo Lycke, Nora Siupsinskiene; Voice Range Profiles of Singing Students: The Effects of Training Duration and Institution. Folia Phoniatr Logop 28 October 2016; 68 (2): 53–59.

[16] Nallanthighal, V. S., A. Härmä, and H. Strik. 2020, May. Speech breathing estimation using deep learning methods. ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 1140–44. IEEE.

[17] J. Wilkins, Prem Seetharaman, Alison Wahland Bryan Pardo, "VocalSet: A Singing Voice Dataset". Zenodo, Mar. 08, 2018. doi: 10.5281/zenodo.1442513.

Appendix A: First Project Review

Title: Listen2Me: Using voice recognition for voice video classification

Student: Kaiqiang Huang

Description (brief): A web application that classifies videos based on their audio. The audio is extracted from the videos and analysed after which it is categorised into a category.

What is complex in this project: Dealing with noise in videos or videos where the is no human speech would be hard to classify.

What technical architecture was used: Tomcat server with Mysql as the database. For voice recognition IFlyTek was used.

Explain key strengths and weaknesses of this project, as you see it: I think the project had a very major weakness from the start, using only audio for video classification. If the student had used image classification combined with audio, better classification could be achieved.

Appendix B: Second Project Review

Title: Artificial Intelligent Voice Assistant

Student: Angesh Kumar Chanderdip

Description (brief): VOYA is a voice assistant app that can listen for certain voice commands and perform functions accordingly.

What is complex in this project: Using NLP to understand the user's commands.

What technical architecture was used: PyQt5.

Explain key strengths and weaknesses of this project, as you see it: The project is very interesting and has a solid use. However, it is challenging to create a fully functional voice assistant.