A Comparison of Fitness Functions in a Genetic Algorithm for Acoustic–Articulatory Parameter Inversion of Vowels

Extended Abstract

Jared Drayton
Interdisciplinary Centre for Computer Music Research
Plymouth University
Plymouth, Devon PL4 8AA, United Kingdom
jared.drayton@plymouth.ac.uk

Eduardo Miranda
Interdisciplinary Centre for Computer Music Research
Plymouth University
Plymouth, Devon PL4 8AA, United Kingdom
eduardo.miranda@plymouth.ac.uk

Alexis Kirke
Interdisciplinary Centre for Computer Music Research
Plymouth University
Plymouth, Devon PL4 8AA, United Kingdom
alexis.kirke@plymouth.ac.uk

ABSTRACT

Articulatory speech synthesis provides an alternative to the state of the art concatenative and formant systems, holding potential for more versatile and expressive artificial speech due to its physical modelling basis. However, a major limitation of practical articulatory synthesis is gaining adequate control of the complex underlying physical models, which stems from a lack of articulatory data. In an effort to procure more data, a Genetic Algorithm approach to Acoustic-Articulatory Parameter Inversion is taken. This paper presents the initial results from testing a number of fitness functions for the Acoustic-Articulatory Parameter Inversion of three vowels, /a/, /o/, and /e/. Three feature vector representations of the vowels were tested; Hertz, Mel–scale, and Cents, in conjunction with three distance metrics. The distance metrics defined the fitness score by calculating the similarity between a candidate and targets feature vector. A Voiced/Un–Voiced constraint was also added as a penalty function, and an indicator of loudness was implemented using a Root Mean Square based co-efficient. The results indicated that certain combinations of the above could lead to convergence towards all three vowels. However, the quality of convergence was not uniform.

CCS CONCEPTS

•**Information systems** →**Speech / audio search;** •**Computing methodologies** →*Speech recognition;*

KEYWORDS

Genetic Algorithm, Speech Synthesis, Physical Modelling

ACM Reference format:
Jared Drayton, Eduardo Miranda, and Alexis Kirke. 2017. A Comparison of Fitness Functions in a Genetic Algorithm for Acoustic–Articulatory Parameter Inversion of Vowels. In *Proceedings of GECCO '17 Companion, Berlin, Germany, July 15-19, 2017,* 2 pages.
DOI: http://dx.doi.org/10.1145/3067695.3076112

1 INTRODUCTION

Existing applications of speech synthesis are dominated by two distinct methods, concatenative [3] and formant [4]. An alternative method, articulatory synthesis, produces artificial speech by computationally modelling the physical phenomena displayed by the human vocal apparatus during speech production. These models are controlled by specifying parameter values that describe a vocal tract area function using a articulation model, and other physical properties. The physical modelling nature presents several opportunities for articulatory synthesis to surpass the state of the art methods, some of these are outlined by Shadle [7]. These include the ability to alter the underlying anatomical aspects of the model in order to recreate an entirely different speaker, and improved co–articulatory behaviour as parameters may be interpolated or transitioned between in a way analogous to actual speech production.

Despite these and other possible benefits, there is a distinct lack of commercial systems employing articulatory synthesis. Several factors contribute to this absence, the largest being the difficulty in controlling such synthesisers, as to measure articulatory actions directly requires specialist equipment. This can be costly and impractical to collect on a large scale when using methods such as Magnetic Resonance Imaging [8]. Therefore, strong motivations exist for developing approaches to obtaining articulatory data from only the speech signal. Estimating articulatory information from a given speech signal is known as "Acoustic–Articulatory Parameter Inversion". A survey of such techniques can be by Schroeter [6]. This papers focus is to investigate the application of a genetic Algorithm in an Analysis by Synthesis approach to Acoustic–Articulatory Parameter Inversion. This is achieved by designing and comparing the performance of several fitness functions in a canonical Genetic Algorithm, tasked with finding parameters for recreating three given target vowels using the PRAAT articulatory synthesiser.

Acoustic–Articulatory Parameter Inversion by Genetic Algorithm has been previously attempted. For example, McGowan [5] harnessed a GA to recover the Task–Dynamics of the ASY synthesiser with promising results. A more recent endeavour, using a multi–population GA approach, was undertaken by Brito [2]. This paper builds on the previous work by using a more complex physiologically informed mass–spring model called PRAAT, with a total of 29 parameters, developed by Boersma and Weenink [1].

GECCO '17 Companion, Berlin, Germany
978-1-4503-4939-0/17/07...$15.00
DOI: http://dx.doi.org/10.1145/3067695.3076112

Additionally, the vocal folds are modelled in the same way as the vocal tract allowing for sub and supra–glottal interaction.

2 METHODS

A formalisation of the problem can be given as (S, f), where S is the set of all valid parameter inputs, and f is an objective function that assigns a value representing similarity between a candidate and target sound. All input parameters p_i may only take values in the interval $-1 \leq p \leq 1$, and are rounded to one decimal place. Therefore, the size of the search space is equal to 21^n, where n is the number of parameters to be optimised.

Three vowels were used as targets for Acoustic–Articulatory Parameter inversion, /a/, /o/, and /e/. They were selected for their contrasting characteristics both perceptually, and in the frequency domain. Similar to [2], the features of interest are formant frequencies, which were extracted using the Linear Predictive Coding (Burg method) built into the PRAAT software. Only the first two formants are used in the feature vectors. In addition to having a list of formants measured in Hz, two more representations were formulated. The first converts the frequencies to the Mel scale, which attempts to compensate for the frequency dependant differences in the perception of loudness. The second is Cents, which is a measure of the difference between two frequencies based on the equal temperament scale. To compare the two feature vectors (x, y), three different distance metrics were used that represented the similarity to the target vowels i.e. the candidates fitness, the Sum of Absolute Difference, the Sum of the Squared Difference, and Euclidean Difference. This resulted in nine different combinations of representation and distance metric.

The above was extended by adding a VUV (Voiced/Un–voiced) penalty and a Root Mean Square co-efficient. The penalty was applied to sounds which had no periodic oscillation of the vocal folds, and assigned a value of 11000 for each formant value. The co-efficient was used to take into account differences in the perceived "loudness" of the sounds. This used the Root Mean Square (RMS) of each signals sample values and would multiply the fitness value i.e. a larger discrepancy between two sounds will result in a greater multiplier.

Genetic operators were assigned the following values for all tests, Population Size - 75 Generations - 20, Crossover - One Point Crossover, Selection - Fitness Proportional, Mutation Probability - 0.15, Gaussian Standard Deviation - 0.15. A real value encoding was used, with two PRAAT parameters being predefined, the Lungs and Levator Palatini, along with The Lungs were predefined due to their time varying nature, and the Levator Palatini was set to 1.0 for the duration to simulate closing off the nasal cavity.

3 RESULTS

Due to the large number of sounds produced, a selection are provided to highlight various behaviours exhibited by the GA. The sound file can be found at the following web address https://soundcloud.com/jareddrayton/sets/gecco-2017. Each file plays the target vowel and then the best rated individual from each generation sequentially for each of the three vowels using the tests using the SSD distance metric and Cents feature representation. In Example 1, a high number of aspirated and non vocalised sounds are present, this was due to PRAAT falsely returning formant frequencies when there were none present. Example 2 illustrates how the VUV penalty greatly lowered the number of sounds which had falsely identified formant values, leading the Genetic Algorithm to exploit better areas of the search space. However, some sounds were particularly quiet with a breathy and falsetto like timbre. Example 3 makes use of the RMS Co–efficient, and demonstrates the ability for convergence towards all three of the vowels and had an increase in the quality of the vowels. Some more general observations from more informal listening tests included the following.

- Out of the three feature representations, Mel consistently failed to perform as well as the Cents and Hz.
- From the distance metrics, SSD and SAD outperformed EUC.
- The /e/ vowel appeared to be the hardest for the Genetic Algorithm to converge on, regardless of the fitness function used.

4 CONCLUSION AND FUTURE WORK

In summary, certain implementations of the fitness functions presented have shown an ability to guide a canonical GA toward converging on three perceptually distinct vowels, albeit to varying degrees of success. The PRAAT model shows a propensity for producing vowels similar to /a/ in the informal listening tests and this is consistent with the more accurate mean and consistently lower standard deviation values for F1 and F2. Immediate future work will be the study of different genetic operator values, additional operators such as elitism and linear ranking, increased number of formants, pitch, and weightings. Following this, the augmentation and development of new fitness functions using different signal processing techniques should be pursued to account for sounds such as fricatives and consonants.

REFERENCES

[1] Paul Boersma and Vincent van Heuven. 2001. Speak and unSpeak with PRAAT. *Glot International* 5, 9-10 (2001), 341–347.

[2] Jose Brito. 2007. Genetic learning of vocal tract area functions for articulatory synthesis of Spanish vowels. *Applied Soft Computing* 7, 3 (jun 2007), 1035–1043. https://doi.org/10.1016/j.asoc.2006.05.004

[3] Andrew J Hunt and Alan W Black. 1996. Unit selection in a concatenative speech synthesis system using a large speech database. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, Vol. 1. 373–376. https://doi.org/10.1109/ICASSP.1996.541110

[4] Dennis H Klatt. 1980. Software for a cascade/parallel formant synthesizer. *The Journal of the Acoustical Society of America* 67, 3 (1980), 971–995.

[5] Richard S McGowan. 1993. Implementing a Genetic Algorithm to Recover Task-dynamic Parameters of an Articulatory Speech Synthesizer. *Haskins Laboratories Status Report on Speech Research SR-113* (1993), 95–106.

[6] Juergen Schroeter and Man Mohan Sondhi. 1994. Techniques for estimating vocal-tract shapes from the speech signal. *IEEE Transactions on Speech and Audio Processing* 2, 1 (1994), 133–150. http://ieeexplore.ieee.org/xpls/abs

[7] Christine H Shadle and Richard I Damper. 2001. Prospects for articulatory synthesis: A position paper. In *4th ISCA Tutorial and Research Workshop.* http://eprints.soton.ac.uk/256064/2/paper.pdf

[8] Brad H Story, Ingo R Titze, and Eric A Hoffman. 1996. Vocal tract area functions from magnetic resonance imaging. *The Journal of the Acoustical Society of America* 100, 1 (1996), 537–554.