A Genetic Approach for Recovering Vocal Tract Area Functions from Spanish Vowels

JOSÉ BRITO WLADIMIR RODRÍGUEZ
Postgrado en Computación
Universidad de Los Andes
Mérida 5101
VENEZUELA

Abstract: This paper shows how a genetic algorithm is able to recover vocal tract area functions from natural utterances. The kind of data analyzed is a subset of Spanish speech signals, concretely vowels from Venezuelan SpeechDat database of utterances, increasing novelty of the study. The method evolves parametric and real-coded representations of speech articulators, with the goal set to minimizing acoustic distance respect to the target, natural SpeechDat utterances. This distance is based on signal's formants and a measure of continuity of the area function. Furthermore, the genetic algorithm is implemented by using the multipopulation approach, seeking to accelerate convergence to a solution while keeping genetic diversity. Subsequently, best learned functions are provided as input to an articulatory speech synthesizer, in order to generate artificial utterances, potentially and acoustically similar to the natural signals. Objective and subjective tests on these artificial signals have positively verified effectiveness of the genetic approach.

Key-Words: Genetic algorithms, Articulatory speech synthesis, SpeechDat, Acoustical models, Inverse speech mapping.

1 Introduction

Articulatory speech synthesis by computer relies on a physical model of speech production, trying to mimic, within computational and knowledge bounds, the *human phonatory system*. Modeling the phonatory system is a hard task, because this system comprises several physiologically complex components, some of which are depicted in Figure 1. The precise set of parameters an articulatory synthesizer requires depends on the underlying physical model. However, independently of the realized model, the hardest part of articulatory systems corresponds to determination of optimum parameters for the synthesizer, in order to produce understandable, natural sounding utterances. Furthermore, the parameters will always be related to temporal behavior of the speech organs during articulations. Particularly, this research uses the Vocal Tract Area Function as a parametric representation of articulators, and as input to a synthesizer, whose structure will be carefully reviewed. Moreover, instead of manually coding the area functions from experimental and human measures, the goal is to develop a genetic algorithm approach to learn the functions from the natural signals. These natural or *target* signals are selected from a vowels subset of Venezuelan SpeechDat database of utterances [13].

This work belongs to the class of inverse problems collectively known as acoustic-to-articulatory mapping, where the goal is to acquire model parameters for a specific synthesizer from direct or indirect analysis of speech signals [3, 15]. The difficulty, however, is that the mapping is nonlinear and non-unique [20]. Over the years, several groups have investigated this problem. For example, Yehia and Itakura adopted an approach based on geometric representations of the articulatory space, including spatial constraints [24]. Dusan and Deng used analytical methods to recover the vocal tract configurations [5]. Sondhi and Schroeter relied on a codebook technique [20]. Genetic algorithms have also been used, albeit the approach and type of signals studied differ to those used in this research [12, 19]. These later studies mainly investigate relations between articulation and perception on the basis of the *tasks* of the task-dynamic description of inputs to a synthesizer [9]. In general, this is the first research applying multipopulation real-coded genetic algorithms to retrieve parametric articulatory representations of Venezuelan Spanish signals.

2 Basic Anatomy and Physiology of Speech

A very partial depiction of speech organs is presented in Figure 1, and they are naturally grouped into three systems [1, 21]:

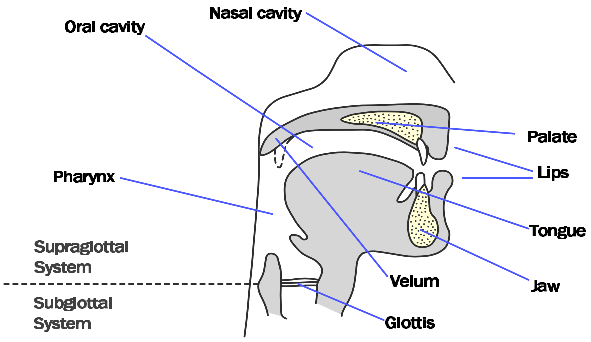

Figure 1: Partial and midsagittal view of phonatory system.

- **Subglottal System**: It is the system below the glottis, and consists of the organs responsible for providing and transporting the excitation energy of the phonatory system, namely, lungs, bronchi, trachea, diaphragm and some other muscles.
- **Laryngeal System**: It refers to the larynx or *voice box* and its surrounding structures. Undoubtedly, the most important of such structures are the vocal folds, which act as modulators of the energy emitted by the subglottal system.
- **Supraglottal System**: These are the articulators and airways above the larynx. They will additionally alter the spectrum of the forward signal, transforming it to an intended or target spectral distribution, according to the linguistic message that is to be transmitted. In short, palate, velum, jaw, tongue, mouth walls, teeth and lips, can modify the sonorous emission in its propagation toward the exterior, across the vocal and nasal tracts. Now, the entire vocal tract may be viewed as a nonuniform acoustic tube running from the glottis to the lips [4, 21].

3 Source-Filter Model of Speech Production

For modeling the complex phonatory system it is required to be aware that a complete theory of speech production is not yet available [16]. Furthermore, operation and interaction of most articulators is highly nonlinear [2, 18, 17], and consequently, for keeping analytical tractability of the problem, linearity assumptions are frequently made. By far the most popular model for digitally working with speech production is the Source-Filter model [6], which is based on a hypothesis of phonatory system dynamics being linear and, specifically, separable into two building blocks: a glottal energy (source) and the vocal tract (filter). This way, voice results from filtering or spectral variations induced by vocal tract configurations on the excitation signal. The glottal source roughly matches the subglottal and laryngeal systems previously presented, while vocal tract (in the following, VT) corresponds to the supraglottal system.

As told, the role of articulators amounts to altering the shape of the supraglottal airways, consequently filtering out or reinforcing some spectral components. As target signals for this study are non-nasalized and isolated spanish vowels, the only airway of interest is the VT. The *Vocal Tract Area Function* is a parametric representation of VT configurations. It shows the cross-sectional area of the VT as a function of distance. The Vocal Tract Area Function is denoted by $A(x)$, where x runs from 0 at the glottis to some specified length L up to the airway opening in the mouth. Typical length from glottis to mouth for the VT of an adult male measures about 17.5 cm [4]. The problem, then, is to recover area functions automatically from target speech signals. Other approaches are to *manually* define $A(x)$ or interpolate images captured in x-ray films or by magnetic resonance. Both of these approaches, however, require considerable effort and domain data.

4 Acoustical Model for Synthesis

The transmission line approach is used in this study to produce artificial utterances from a given vocal tract area function. For a lossless uniform cross-sectional area tube, an analogy is established with electrical wave propagation along transmission lines, where for a voltage $e(x,t)$ and current $i(x,t)$, the following relations hold [6, 7]:

$$\begin{aligned} -\frac{\partial e}{\partial x} &= L\frac{\partial i}{\partial t} \\ -\frac{\partial i}{\partial x} &= C\frac{\partial e}{\partial x} \end{aligned} \tag{1}$$

L and C are inductance and capacitance per unit length of the transmission line. Accordingly, $L = \rho/A$ and $C = A/\rho c^2$ are referred to as *acoustic inductance* and *acoustic capacitance*, respectively. In turn, $\rho \approx 1.14 \times 10^{-3}\ gm/cm^3$ is the air density, and $c \approx 344\ m/s$ is the sound velocity. In order to model non-uniform cross-sectional area configurations, the VT will be represented by several abutting *tubelets* of constant cross-sectional area, discretizing $A(x)$. As each tube can be modeled by a quadripole network, a series of such elements is used to form the entire computational VT. Losses are incorporated to the model as depicted in Figure 2. There, elements R and G stand for acoustic losses due to viscous friction and heat conduction, respectively. Complete physical interpretations are given by Flanagan [7].

By using impedances, the network is simplified to elements Z_1 and Z_2, whose expressions are [6]:

$$\begin{aligned} Z_1 &= Z \tanh \lambda/2 \\ Z_2 &= Z/\sinh \lambda \\ Z &= \sqrt{\frac{R + j\omega L}{G + j\omega C}} \\ \lambda &= l\sqrt{(R + j\omega L)(G + j\omega C)} \end{aligned} \tag{2}$$

During generation of artificial utterances, frequency ω is fixed to a predefined value. However, the genetic algorithm must also recover l_i, the length for each tubelet i approximating VT. In turn, A_i is the respective cross-sectional area. Then, for all i, A_i and l_i represent a VT configuration. Up to this point, discussion has centered on supraglottal modeling. In order to produce an acoustic output, it is also required to define the glottal source U_g, which is modeled as a current source. In this study, a modal source is generated by means of the Liljencrants-Fant (LF) model [10, 21], because of its high quality and acceptance. Furthermore, a simplification is made by restricting target signals to be from male, low-pitched speakers, which allows to set F0 to a proper and fixed frequency of 125Hz. Finally, the model also includes Flanagan's terms for modeling radiation at the lips [7].

5 Recovering Tubelets Specifications

A multipopulation Genetic Algorithm [22, 23] will be used for recovering VT configurations from the irregularly shaped solution space. This model recurs to an ensemble of subpopulations (islands), instead of a single population. In every generation, the GA proceeds on each subpopulation in an independent fashion. However, every fixed number of generations, a special *migration* operator moves individuals from one subpopulation to another. This multipopulation approach helps to keep genetic diversity. Therefore, if a subpopulation converges prematurely to a suboptimal or anatomically invalid VT configuration, genetic material proceeding from other populations may redirect the search toward a better solution.

5.1 Representation of Population and Genetic Operators

$A(x)$ is discretized according to the division of the VT in S_{max} sections, implying that $1 \leq i \leq S_{max}$. In other words, S_{max} tubelets are going to be used for modeling the VT. Consequently, as the length l_i of each tubelet must also be considered, the phenotype will consist of 2 × S_{max} real values. Range for cross-sectional areas is restricted to $[0.0, 9.9]\ cm^2$. On its side, l_i values will lie in $[0.0, 1.5]\ cm$. These bounds drive out some unnatural solutions. By using a real-valued encoding [8], genotype is one-to-one mapped, exhibiting the composition in Figure 3.

In subsequent experiments, 12 islands are used, each hosting 20 individuals, so yielding a per generation total of 240 individuals. The GA is run until 1000 generations are created. Also, because the genotype is based on real encoding, *discrete recombination* (uniform crossover with real valued alleles) is used for breeding of offspring. On the other hand, selection applies classic *stochastic universal sampling*. And with respect to mutation, a real-tailored operator with

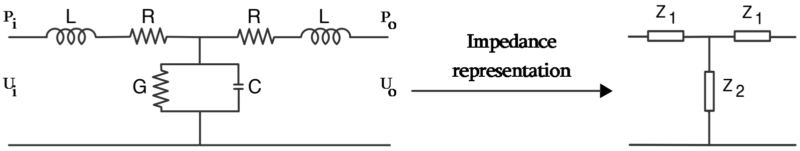

Figure 2: Quadripole network accounting for energy losses.

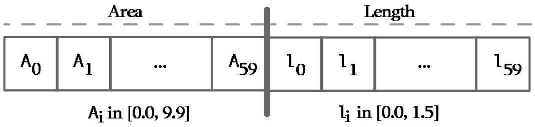

Figure 3: Genotype structure with real alleles.

10^{-3} probability of mutation is used. Besides, reinsertion of individuals in a population follows a simple *fitness-based* approach. Finally, migration operator is assigned a 0.3 probability, with a *neighborhood strategy*, which implies that migration of individuals only occurs between adjacent populations.

5.2 Objective Function

The objective function F allows to discern *good* VT configurations. Concretely, the goal is to minimize acoustical distance between artificial and natural, target utterances. In other words, given a fixed $U_g(t)$, the best individuals are those defining a VT configuration whose acoustic output is very near to the target signal, in the sense of the distance criterion. In this study, such acoustic distance is represented by *formant* distance. Formants are resonances of VT, and by minimizing distance, it is expected to obtain the right parameters for an artificial VT matching natural resonances. In formal terms, the goal is to minimize F, where

$$\mathtt{F} = formantDistance(A, l, TargetSignal) \quad (3)$$

A and l are the area and length components of the VT configuration. $formantDistance$ means weighted distance between *vector of predicted formants* and *vector of target formants*. However, a metric solely based on acoustic distance does not impose constraints on the shape of the artificial VT, i.e., derived A_i and l_i, although acoustically good, may correspond to a configuration unrealizable by the human phonatory system. Heuristically, for precluding anatomically unfeasible configurations, F is extended with two $shapeFactor$ terms penalizing discontinuity of A_i and l_i sequences along index i. This way, evolution is guided to prefer smoother VT configurations. Accordingly, F is now stated as:

$$\begin{aligned}\mathtt{F} = {} & formantDistance(A, l, TargetSignal) + \\ & + shapeFactor(A) + shapeFactor(l)\end{aligned} \quad (4)$$

$shapeFactor(A)$ is simply

$$\sum_{j=2}^{S_{max}} \frac{|A_j - A_{j-1}|}{A_j}$$

A similar relation applies to $shapeFactor(l)$. On its side, $formantDistance$ requires computing the two formant vectors, one for formants associated with a given VT configuration (V_P), and another for the target speech signal (V_T).

5.2.1 Computing vector of predicted formants

With respect to our acoustical model, formants are the poles of the acoustic transfer function $H(\omega)$ determined by a specific VT configuration. $H(\omega)$ is computed as the relation between acoustical output and input of the overall system, formally, $H(\omega) = U_r/U_g$. Once $H(\omega)$ has been computed, all that remains is to apply the N_b method [10], in order to approximate its peak frequencies. This way, first four formants for the given VT are computed and arranged in vector V_P.

5.2.2 Computing vector of target formants

In order to track the first four formants of a target signal, Linear Prediction (LP) analysis is applied. This way, an all-pole filter is derived as a model for the VT which produced the utterance. Subsequently, filter's poles are interpreted as the sought resonances, and grouped into the vector of target formants V_T. Now, let $d(k) = |V_T(k) - V_P(k)|/V_T(k)$, where $1 \leq k \leq 4$ indexes formants F1-F4 in vectors V_T and V_P. Finally, $formantDistance$ is defined as $formantDistance = 40d(1) + 30d(2) + 15d(3) + 10d(4)$, weighting formants according to their importance in vowels' formation.

6 Experiments and Results

In this section the experiment is detailed and its results presented. The first step is to form the corpus of target signals. This corpus includes 25 vowels (five instances of each spanish vowel, `/a/`, `/e/`, `/i/`, `/o/` and `/u/`) of different low-pitched male speakers, selected from SpeechDat database of Venezuelan Spanish utterances [13]. For each target signal in the corpus, the multipopulation GA was applied to recover best VT configurations, with S_{max} set to 60. The multipopulation approach was distributedly implemented in three desktop workstations, one of which was appointed as the central node. During any generation, each workstation was loaded evaluating four populations, which as expected largely reduced the total computation time with respect to preliminary single population essays.

In order to assess quality of synthetic utterances, a group of 8 Venezuelan evaluators was formed for perceptual tests. For each evaluator, a software routine shuffled artificial signals and played them. After playing a signal, the program asked the listener to input the vowel he/she had heard. An error was counted by the program if the user had input a vowel distinct to the played one. Evaluators were not aware of the proportion of vowels in the test corpus. Table 1 presents errors as ratios: not recognized instances of a vowel divided by total instances of that vowel. In this case, the global mean error of 0.41 is an acceptable rate, considering that F0 is fixed and only 1000 generations were run for approximating each target signal.

Additionally, the GA was able to derive anatomically consistent VT configurations, respect to shape of the area functions. That is a highly satisfactory result, considering the simple heuristics embodied in F, and the lack of measures about the natural VT (except by signals themselves). For example, area function for `/a/` corresponded to an open VT. Vowel `/o/`, although fairly open, exhibited certain closure in the posterior VT. `/u/` also has a posterior, while more pronounced closure. On its side, `/e/` and `/i/` had palatal area reductions, with `/i/`'s reduction lower that `/e/`'s, which is the only incorrect result. In fact, Table 1 confirms that `/i/` was the vowel with lowest approximation quality. Still, the pattern of the area function for `/i/` is not too far from the right configuration. All in all, these configurations are consistent with the articulatory traits of spanish vowels [11, 14].

	Vowels					
Evaluators	/a/	/e/	/i/	/o/	/u/	Mean
Eval1	0.20	0.20	0.60	0.40	0.30	0.34
Eval2	0.00	0.40	0.60	0.60	0.60	0.44
Eval3	0.20	0.20	0.60	0.60	0.30	0.38
Eval4	0.20	0.60	0.60	0.60	0.60	0.52
Eval5	0.20	0.20	0.80	0.80	0.60	0.52
Eval6	0.20	0.00	0.60	0.30	0.30	0.28
Eval7	0.00	0.20	0.80	0.30	0.60	0.38
Eval8	0.40	0.20	0.60	0.30	0.40	0.38
	0.18	0.25	0.65	0.49	0.46	0.41

Table 1: Ratio of errors in recognition of artificial vowels by evaluators.

7 Conclusions

Both objective and subjective evaluations positively verify the feasibility of the developed multipopulation GA for retrieving the parametric model underlying natural utterances, concretely male vowels selected from the Venezuelan SpeechDat. A recognition error lower than 0.5 is fairly good, considering that F0 was fixed although all the target signals did exhibit differences in fundamental frequency. Then, further research is needed accounting for pitch traits in target signals, in order to approach signals of diverse frequency, not only male voices. An additional attractive of the developed approach is the simplicity of the objective function, combining formant distance with shape factors for penalizing solutions with very abrupt area variations. Furthermore, only 1000 generations and 12 islands per essay were used; increasing those bounds may lead to better VT configurations.

References

[1] Rodney Ball. Introduction to phonetics for students of english, french, german and spanish. http://www.lang.soton.ac.uk/profiles/ball.htm.

[2] José Brito. Identificación de señales verbales en el espacio de fase reconstruido. Master's thesis, Universidad de Los Andes, June 2004.

[3] M. A. Carreira-Perpinán. *Continuous latent variable models for dimensionality reduction and sequential data reconstruction*. PhD thesis, Department of Computer Science, University of Sheffield, Febrero 2001.

[4] P.B. Denes and E.N. Pinson. *The Speech Chain: The Physics and Biology of Spoken Language*. W.H. Freeman and Company, New York, USA, 1993. Second edition.

[5] S. Dusan and L. Deng. Acoustic-to-articulatory inversion using dynamic and phonological constraints, 2000. 5th Speech Production Seminar.

[6] Gunnar Fant. *Acoustic Theory of Speech Production*. Description and Analysis of Contemporary Standard Russian. Mouton, The Hague, The Netherlands, 1970.

[7] James Loton Flanagan. *Speech Analysis, Synthesis, and Perception*. Springer-Verlag, Berlin, 1972. Second edition.

[8] C. Janikow and Z. Michalewicz. An experimental comparison of binary and floating point representations in genetic algorithms. In *Proceedings of the 4th International Conference on Genetic Algorithms*, San Diego, USA, 1991. Morgan Kaufmann.

[9] J. Kelso, E. Saltzman, and B. Tuller. The dynamical perspective on speech production: data and theory. *Journal of Phonetics*, 14:29–59, 1986.

[10] Qiguang Lin. *Speech Production Theory and Articulatory Speech Synthesis*. PhD thesis, Royal Institute of Technology, Stockholm, Sweden, 1990.

[11] Bertil Malmberg. *La fonética*. Editorial Universitaria de Buenos Aires, 1970. Fourth edition.

[12] R. S. McGowan. Recovering articulatory movement from formant frequency trajectories using task dynamics and a genetic algorithm: Preliminary model tests. *Speech Communication*, 14, 1994.

[13] A. Moreno and E. Mora. Speechdat spanish venezuelan database for the fixed telephone network. Technical report, Universidad Politécnica de Cataluña, España y Universidad de Los Andes, Venezuela, 1999.

[14] Enrique Obediente Sosa. *Fonética y Fonología*. Consejo de Publicaciones de la Universidad de Los Andes, 2001. Third edition.

[15] H. Ploner-Bernard. Speech synthesis by articulatory models. Technical report, Graz University of Technology, 2003.

[16] Thomas Quatieri. *Discrete-time speech signal processing*. Prentice-Hall, New Jersey, USA, 2002.

[17] W. Rodríguez, H.-N. Teodorescu, F. Grigoras, A. Kandel, and H. Bunke. A fuzzy information space approach to speech signal non-linear analysis. *International Journal of Intelligent Systems*, 15(4):343–363, 2000.

[18] Wladimir Rodríguez. *Similarity of Dynamical Systems*. PhD thesis, University of South Florida, 1998.

[19] E. Saltzman. Task-dynamic coordination of the speech articulators: A preliminary model. *Experimental Brain Research*, 15, 1986.

[20] J. Schroeter and M. Sondhi. Techniques for estimating vocal-tract shapes from the speech signal. *IEEE Transactions on Speech and Audio Processing*, 2, 1994.

[21] Kenneth N. Stevens. *Acoustic Phonetics*. Current studies in linguistic. The MIT Press, Massachusetts, 1998.

[22] Darrell Whitley. A genetic algorithm tutorial. *Statistics and Computing*, 4:65–85, 1994.

[23] Darrell Whitley and T. Starkweather. Genitor 2: a distributed genetic algorithm. *Journal of Experimental and Theoretical Artificial Intelligence*, 2:189–214, 1990.

[24] H. Yehia and F. Itakura. A method to combine acoustic and morphologial constraints in the speech production inverse problem. *Speech Communication*, 18, 1996.