INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Articulatory Copy Synthesis Based on A Genetic Algorithm

Yingming Gao, Simon Stone, Peter Birkholz

Institute of Acoustics and Speech Communication, TU Dresden, Germany

yingming.gao@mailbox.tu-dresden.de, {simon.stone, peter.birkholz}@tu-dresden.de

Abstract

This paper describes a novel approach for copy synthesis of human speech with the articulatory speech synthesizer VocalTractLab (VTL). For a given natural utterance, an appropriate gestural score (an organized pattern of articulatory movements) was obtained in two steps: initialization and optimization. In the first step, we employed a rule-based method to create an initial gestural score. In the second step, this initial gestural score was optimized by a genetic algorithm such that the cosine distance of acoustic features between the synthetic and natural utterances was minimized. The optimization was regularized by limiting certain gestural score parameters to reasonable values during the analysis-by-synthesis procedure. The experiment results showed that, compared to a baseline coordinate descent algorithm, the genetic algorithm performed better in terms of acoustic distance. In addition, a perceptual experiment was conducted to rate the similarity between the optimized synthetic speech and the original human speech. Here, similarity scores of optimized utterances with regularization were significantly higher than those without regularization.

Index Terms: articulatory copy synthesis, acoustic-to-articulatory inversion, genetic algorithm, parameter regularization

1. Introduction

Compared to unit-selection synthesis or statistical parametric speech synthesis, articulatory synthesis can provide flexible controls in speech generation and speech visualization [1]. It benefits speech training of speech-impaired or hearing-loss people, foreign language learning of normal-hearing students, and speech production and perception research [2, 3]. Articulatory synthesis takes a time series of a set of parameters describing a vocal tract model and converts it to an acoustic speech signal using an acoustic model. The parameter time series are usually obtained according to one of two paradigms: (1) imitate the assumed articulation underlying a speech sound as closely as possible until some subjective quality criterion is met, or (2) use a natural utterance as a reference and vary the articulatory parameters until the resulting speech signal is as similar as possible to the reference, i.e., copy synthesis.

There are two types of methods to determine articulatory parameters for copy synthesis. One of them is to make articulatory recordings, e.g., using electropalatography (EPG), X-ray microbeam, electromagnetic articulography (EMA), or magnetic resonance imaging (MRI). Vocal tract shapes are measured from articulatory data and used to estimate area functions for speech synthesis [4–6]. Another is to derive articulatory parameters directly from speech signals (i.e., acoustic-to-articulation inversion). It is inexpensive and non-invasive to collect speech signals compared to articulatory data. Dang and Honda [7] implemented the estimation of tongue-position control points from vowel formants in their physiological articulatory model. They used analysis-by-synthesis (AbS) to minimize the formant distance between the input sound and the synthetic sound, and finally, reproduced Japanese vowel-to-vowel sequences.

When acoustic-to-articulation inversion is applied to connected speech, the articulatory movements can be represented in terms of a gestural score, a concept stemming from articulatory phonology [8]. A gestural score is an organized pattern of multiple articulatory gestures for the realization of an utterance. In general, a gesture represents movement toward a target configuration of the vocal tract model or the vocal folds by the participating articulators/parameters. In the context of articulatory copy synthesis, an appropriate gestural score can be obtained by temporally adjusting gestures to make the synthetic speech sound like the original speech as much as possible, just like a duplicate. Bauer et al. [9] implemented articulatory speech copy synthesis by manually aligning and coordinating phonological gestures with reference to acoustic landmarks. The resulting gestural score was fed into VocalTractLab [10] to reproduce the natural utterance. This process is obviously labor-intensive and subjective. Aiming at facilitating the training of acoustic models for automatic speech recognition (ASR), some researchers developed automatic methods to derive gestural scores from speech signals [11, 12]. However, they made strong assumption that the dynamical parameter (speaking effort) of gestures do not vary from instance to instance, and only durations and relative timings of gestures are allowed to change.

The genetic algorithm has been demonstrated to be effective for real parameter optimization, especially for non-differentiable problems [13]. In the present paper we proposed an approach for automatic articulatory copy synthesis based on a genetic algorithm. This research was conducted with the articulatory speech synthesizer VocalTractLab [1]. The gestural score was represented as a chromosome, and the parameters of the gestures were encoded as genes [13, 14]. A population of individuals (candidate gestural scores) evolved under the law of survival of the fittest. The fittest gestural score corresponded to the synthetic utterance with the least acoustic distance to the input utterance. The final gestural score with best fitness was selected as the solution.

2. Method

2.1. Articulatory speech synthesis with VocalTractLab

VocalTractLab (VTL) is an articulatory speech synthesizer, which simulates the articulation process, specified by gestural scores, and simultaneously produces acoustic signals. As shown in the upper panel of Figure 1, a gestural score is organized in eight tiers corresponding to supraglottal articulation, glottal settings, and lung pressure. The realization of each phoneme is considered to comprise multiple gestures, which are coordinated and distributed over the tiers. Each gesture consists of three parameters [10]: a gesture *value*, a *duration*, and a *time*

http://dx.doi.org/10.21437/Interspeech.2019-1334

constant, which define target positions of articulators, their duration, and how quickly the participating articulators reach the targets (i.e., speaking effort), respectively.

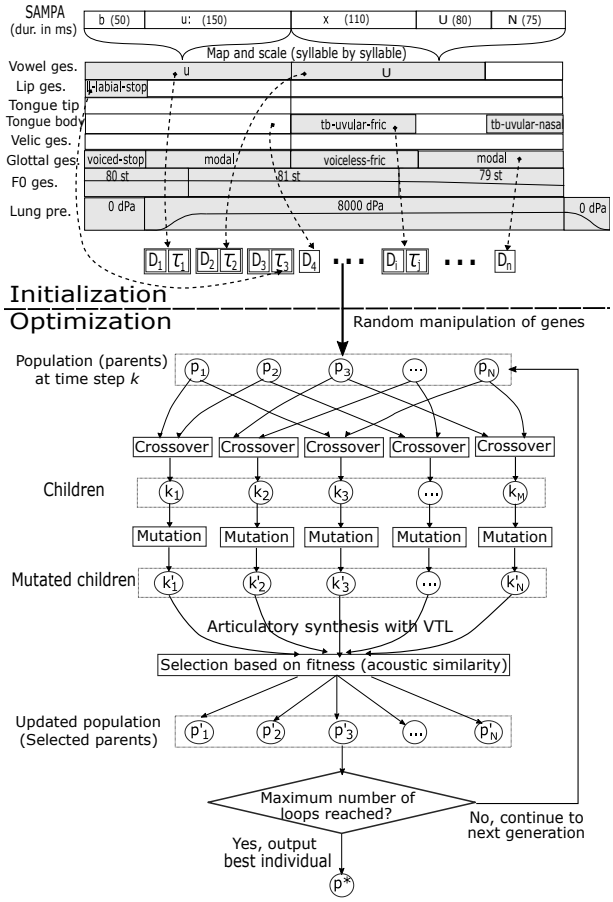

Figure 1: *Schematic diagram of the proposed automatic articulatory copy synthesis based on a genetic algorithm. The upper panel demonstrates the initialization step where phonetic transcriptions in SAMPA are first mapped to gestures and then the durations are aligned with the reference utterance. The lower panel depicts the optimization step.*

2.2. Gestural score initialization

In this step, we employed a rule-based method to create a gestural score from a phonetic transcription of the reference utterance. Each phone of a syllable was first mapped to a supraglottal gesture and a glottal gesture. We adopted the time structure model of the syllable to organize all gestures involved in a word [15]. The temporal alignment of all the phones within a syllable follows some simple principles: the initial consonant and the vowel share the same onset of the syllable, and the other phones are sequentially aligned after the vowel of the syllable. Accordingly, a gestural score for a word is organized as follows (here, for the German word "Buchung" as illustrated in Figure 1): the initial consonantal gesture (like gesture "ll-labial-stop" for consonant [b]) and the vocalic gesture (like gesture "u" for vowel [u:]) start at the syllable onset; the gestures of others phones (like gesture "tb-velar-nas" for nasal [N]) are sequentially arranged after the offset of the vowel gesture. The gesture *durations* were specified with intrinsic durations of German phonemes measured by Kohler [16]. The *time constants* of supraglottal gestures were set to their preferred values [17] (see initial values in second column of Table 2) . The *time constant* of a vowel followed that of its preceding consonantal gesture. We linearly scaled the durations of all gestures within each syllable such that the synthetic and the reference utterance had the same length. In the f_0 tier, the gestures were estimated by the TargetOptimizer [18], and were kept fixed during optimization.

2.3. Optimization based on coordinate descent

Although the synthetic speech of the initialized gestural scores was intelligible, we aimed to reproduce the reference utterance. Therefore, the initialized gestural scores were further fed into the second step for optimization. We implemented a baseline optimization system based on a coordinate descent algorithm [19]. In each time step, we randomly selected a coordinate direction (a parameter of gestures) and minimized the acoustic distance (see Sec. 2.6) by randomly searching neighboring points, while fixing all other coordinates (other parameters to be optimized). In each iteration, every coordinate was selected once in random order. This procedure was repeated until the maximum number of iterations was reached.

2.4. Optimization based on a genetic algorithm

The coordinate descent algorithm may suffer from difficulties in convergence and lack of concurrency, so we proposed another optimization approach based on a genetic algorithm (see the lower panel in Figure 1), which imitates the biological mechanism of evolution. A genetic algorithm starts with an initial population of individuals, where each individual is encoded by its chromosomes. It then produces a new population (the children) by recombining the chromosomes of individuals from the initial population (the parents). The chromosomes of the children then mutate and another generation can be created using the mutated children as parents. In this study, the chromosome represented the gestural score. Each parameter of gestures (*duration* or *time constant* in Sec. 2.1) was encoded as a gene with a real-valued encoding technique [13, 14]. The *initial population* of a certain size was generated by adding normally distributed random values to parameters of the initialized gestural score. Then the chromosomes of *initial parents* were randomly recombined with one another to create *children* by the "crossover" operation, and these children would mutate by randomly adding a random value (see details in Sec. 2.6). Next, all *mutated children* (gestural scores) were fed into VTL to synthesize speech. Those children with good fitness (closer acoustic distance to the reference utterance) were selected and served as the *population* of the next generation. The population evolved until the maximum number of iterations was reached. The aim of the evolutionary strategy was to produce increasingly better individuals over time, so eventually, the final optimal child corresponded to the best gestural score for the natural utterance.

2.5. Reference utterances

Even though the optimization can be applied to any complex utterances, individual words were used in this paper. The words were selected as follows. First, we selected all two- and three-syllable words from the pronunciation dictionary in [20] and the phonetically balanced BITS corpus [21]. Then, frequently used words of them (with a frequency level higher than 15 in the frequency-based ranking list [22]) were kept. Next, a minimal set of words covering all phonemes at least once were selected based on a modified least-to-most-ordered algorithm [23]. We

repeated this step three times to cover more combinations of different consonants and vowels. Finally, the word list contained 30 two-syllable words and 11 three-syllable words. Speech signals of the selected 41 words, spoken in the carrier sentence "Ich habe ... bestellt" by a male German native speaker, were subsequently recorded in an audio studio environment.

2.6. Optimization settings

During optimization, the mutation operation was individually performed on each parameter x_i ($1 < i < n, n$ is the number of durations or time constants to be optimized) by adding a normally distributed random value with expectation zero and standard deviation σ. The mutated parameter x_i' is defined as:

$$x_i' = x_i + \sigma \cdot N_i(0,1) \tag{1}$$

where the so-called step size σ was assigned with a fixed value to all parameters of gestures of *initial population*, but was adapted in each generation over time. The mutation was then rewritten [24] as:

$$\sigma_i' = \sigma_i \cdot \exp\left(\tau' \cdot N(0,1)\right) + \tau \cdot N_i(0,1)) \tag{2}$$

$$x_i' = x_i + \sigma_i' \cdot N_i(0,1) \tag{3}$$

where $\tau' = \frac{c}{\sqrt{2n}}$ and $\tau = \frac{c}{\sqrt{2\sqrt{n}}}$ are the learning rates ($c = 1$ is a reasonable choice when the number of evolution generations is between 20 and 100) [25]. For the assessment of the fitness, the gestural scores of the mutated children were synthesized with VTL. We used the cosine distance of acoustic features to measure the similarity between synthetic and reference utterances. The smaller the distance, the more similar the synthetic and natural utterances are. The cost function was expressed as:

$$\mathcal{L} = \frac{1}{N}\sum_{i=1}^{N}(1 - \frac{\sum_{j=1}^{M} X_{i,j}\hat{X}_{i,j}}{\sqrt{\sum_{j=1}^{M} X_{i,j}^2}\sqrt{\sum_{j=1}^{M} \hat{X}_{i,j}^2}}) \tag{4}$$

where N is the frame number of reference utterance, and M is the number of acoustic features per frame. $X_{i,j}$ and $\hat{X}_{i,j}$ are the jth features of the ith-frame of the natural and synthetic utterances, respectively.

Due to motor equivalence phenomena in speech production [26], the acoustic-to-articulatory inversion suffers from the problem of non-uniqueness. One of the challenges was that different gestural values and temporal coordination of gestures can result in very similar synthetic signals. As a result, the combinatorial explosion made the search space grow rapidly. Nevertheless, the problem can be alleviated by introducing additional articulatory and phonological constraints [7] [27]. Here, by incorporating deviations of time constants from their *preferred* values into the cost function, we limited gestural score parameters to plausible values:

$$\begin{aligned}\mathcal{L} = \alpha\frac{1}{N}\sum_{i=1}^{N}(1 - \frac{\sum_{j=1}^{M} X_{i,j}\hat{X}_{i,j}}{\sqrt{\sum_{j=1}^{M} X_{i,j}^2}\sqrt{\sum_{j=1}^{M} \hat{X}_{i,j}^2}}) \\ +(1-\alpha)\frac{1}{\beta}\frac{1}{K}\sum_{k=1}^{K}|\hat{\tau}_k - \tau_k|^3\end{aligned} \tag{5}$$

where K is the number of time constants of all supraglottal gestures. $\hat{\tau}_k$ and τ_k are the current value and preferred value of the kth time constant, respectively. Their absolute difference is cubed so as to exaggerate those with relatively bigger deviations, thus giving higher penalty to them. β is a scaling factor to make the two costs have comparable magnitudes. α is a weight to balance the two costs.

We used 42 acoustic features: 13 Mel-Frequency Cepstral Coefficients (MFCC) and 1 voiced/unvoiced probability extracted using STRAIGHT [28] as well as their first and second order derivatives. These features were extracted from a 25-millisecond-length window shifted every 10 milliseconds. Using Bayesian optimization [29], we optimized the hyperparameters by minimizing the sum of acoustic distances of five words. Table 1 lists the search ranges and optimized values of hyperparameters.

Table 1: *Hyperparameter optimization*

Hyperparameter	**Search Range**	**Result**
Size of population	Integer $\in [10, 30]$	14
Number of children	Integer $\in [50, 150]$	145
Initial σ of mutation	Decimal $\in [0.0005, 0.005]$	0.00198
Weight α	Decimal $\in [0.7, 0.99]$	0.9

2.7. Perception experiment

In addition to the measurement of the acoustic distance, we also conducted a perception experiment after optimization. 20 German natives rated the similarity between the optimized synthetic and corresponding reference utterances. Each time a word was first prompted on a computer screen and then the reference utterance and one corresponding synthetic utterance were played to the listeners, with an interval of 0.6 seconds between them. The order of the stimuli pairs was randomized for each subject. They were asked to rate the similarity on a 4-point Likert scale with "1" standing for "very different", "2" for "rather different", "3" for "rather similar", and "4" for "very similar".

3. Results and discussion

3.1. Optimization results

We implemented three optimization systems: (1) coordinate descent with regularization of time constants, (2) genetic algorithm without regularization of time constants, (3) genetic algorithm with regularization of time constants. They performed 100 iterations or generations, respectively. Figure 2 demonstrates an example of a final gestural score and corresponding synthetic speech optimized by the genetic algorithm with regularization.

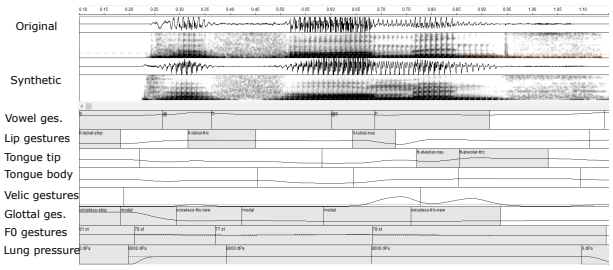

Figure 2: *The optimized gestural score and synthetic speech for word "Performance". Above the optimized gestural score is the oscillogram and spectrogram of synthetic speech. As a reference for comparison, the oscillogram and spectrogram of the reference utterance are displayed above them.*

We compared the performance of the three systems in terms of the acoustic distance, i.e. equation (4). Although the 41 utterances were optimized individually, we summed their acoustic distances after each generations/iteration for the sake of comparison. As we can see from Figure 3, after the first 15 generations/iterations, the acoustic distance remains almost constant for the coordinate descent method while it continues to decrease for the other two methods. The genetic algorithm with regularization outperforms that without regularization a little. However, the benefit of regularization is prominent when we further examine the values of the optimized *time constants*. As we can see form Table 2, the regularized time constants are not only closer to their *preferred* values, but also have smaller standard deviations.

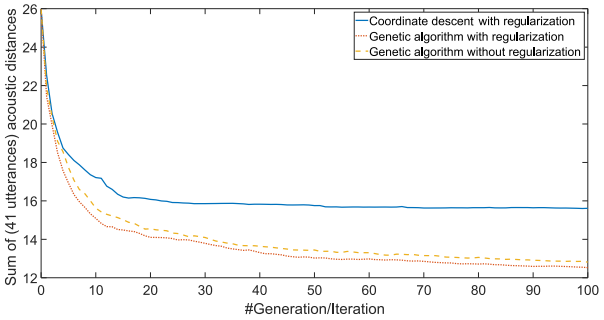

Figure 3: *Acoustic distance (calculated with equation (4)) of different optimization methods after each generation/iteration. The* 0*th generation/iteration indicates the acoustic distance between initialized synthetic speech and original human speech.*

Table 2: *Initial and optimized time constants: average values in milliseconds (standard deviations in parentheses)*

Gestures	**Inital**	**GA-reg.**	**GA-non-reg.**
ll-labial-fricative	10	10.2 (1.37)	13.0 (9.32)
ll-labial-nasal	10	10.7 (1.15)	19.6 (9.05)
ll-labial-stop	10	10.3 (1.06)	12.3 (6.49)
tt-alveolar-fricative	15	14.4 (1.66)	20.0 (12.47)
tt-alveolar-lateral	5	6.3 (1.11)	15.6 (6.69)
tt-alveolar-nasal	15	16.5 (1.16)	23.6 (11.81)
tt-alveolar-stop	15	14.8 (1.55)	16.8 (9.04)
tt-postalveolar-fricative	15	13.3 (1.52)	7.5 (5.12)
tb-palateralal-fricative	20	19.3 (1.34)	18.1 (12.85)
tb-uvular-fricative	15	15.5 (2.09)	23.8 (12.11)
tb-velar-nasal	20	22.3 (0.24)	18.7 (12.48)
tb-velar-stop	15	15.1 (1.94)	12.7 (6.22)

3.2. Perceptual similarity

Figure 4 shows the boxplots of mean similarity scores of pooled words and raters for different methods. Compared to the initialized method (in which gestures were linearly scaled and aligned with reference to natural utterance), the similarity scores increased after optimization. A paired *t*-test shows that, the two optimization methods with regularization achieved significantly higher similarity scores compared to the method without regularization. However, no significant differences were found between the genetic algorithm with regularization and the coordinate descent with regularization, and between the genetic algorithm without regularization and the initial. The perception evaluation seems to be inconsistent with the acoustic measurement. This could be explained that, the acoustic distance was calculated frame-by-frame and the optimization methods sought to the least sum of acoustic distance while listeners rated the similarity based on the overall similarity. Listeners often gave a low score to synthetic speech with a specific unnatural or unintelligible segment but all other segments being well optimized. This may be a reason why the baseline system has a big acoustic distance but a relatively high similarity score.

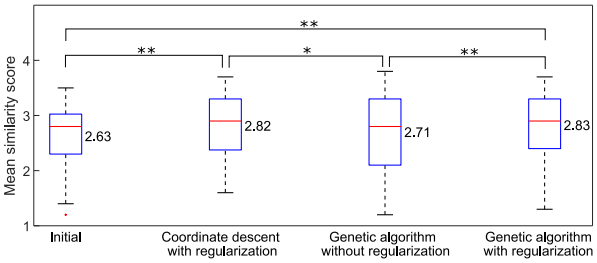

Figure 4: *Boxplots of mean similarity scores for different methods with the scores pooled across words and raters. The numbers next to the boxes indicate the mean values. (Considering Bonferroni correction:* $*p < 0.0083$*;* $**p < 0.0017$*).*

Besides, we found the proposed method did not always work well for all phonemes. For example, nasal consonants sometimes become too weak to perceive, and even totally disappeared after optimization, although they could be clearly perceived in the initial utterances. This may result from the effect of antiformants. Another factor that may influence the optimization is the speaker variation, i.e. the acoustic differences of the same phoneme produced by different speakers, which inevitably distorts the similarity measurement especially when the acoustic distance is calculated frame-by-frame between synthetic and original utterances.

4. Conclusions and future Work

In the present paper, we proposed a novel approach to the automatic copy synthesis of human speech. We employed a rule-based method to create initial gestural scores, and then optimized them using a genetic algorithm, which increased the acoustic similarity between synthetic and reference utterances. By incorporating regularization, we limited the time constants to plausible values, thus obtaining better perceptual similarity scores. Meanwhile, we found that the proposed method did not always work well for all phonemes. In the future, attentions will be focused on seeking robust acoustic features and trying speaker adaptation techniques, e.g. applying vocal tract length normalization (VTLN) to feature extraction. We will also test more words produced by different speakers.

5. Acknowledgements

We would like to thank the participants of the listening experiments. This research work was partially sponsored by China Scholarship Council and German BMW (support code: ZF4443004BZ8).

6. References

[1] P. Birkholz, "Modeling consonant-vowel coarticulation for articulatory speech synthesis," *PloS one*, vol. 8, no. 4, p. e60603, 2013.

[2] H. R. Javkin, E. G. Keate, N. Antonanzas-Barroso, and B. A. Hanson, "Synthesis-based speech training system and method," Jul. 16 1996, uS Patent 5,536,171.

[3] B. Elie and Y. Laprie, "Copy synthesis of running speech based on vocal tract imaging and audio recording," in *ICA 2016 - 22nd International Congress on Acoustics*, 2016.

[4] C. Ericsdotter, "Articulatory copy synthesis: Acoustic performance of an mri and x-ray based framework," in *Proceedings of the XVth ICPhS*, 2003, pp. 2909–2912.

[5] Y. Laprie, M. Loosvelt, S. Maeda, R. Sock, and F. Hirsch, "Articulatory copy synthesis from cine x-ray films," in *INTERSPEECH - 14th Annual Conference of the International Speech Communication Association, August 25-29, Lyon, France, Proceedings*, 2013, pp. 2024–2028.

[6] Y. Laprie, B. Elie, and A. Tsukanova, "2d articulatory velum modeling applied to copy synthesis of sentences containing nasal phonemes," in *International Congress of Phonetic Sciences*, 2015.

[7] J. Dang and K. Honda, "Estimation of vocal tract shapes from speech sounds with a physiological articulatory model," *Journal of Phonetics*, vol. 30, no. 3, pp. 511–532, 2002.

[8] C. P. Browman and L. Goldstein, "Articulatory phonology: An overview," *Phonetica*, vol. 49, no. 3-4, pp. 155–180, 1992.

[9] D. Bauer, J. Kannampuzha, and B. J. Kröger, "Articulatory speech re-synthesis: Profiting from natural acoustic speech data," in *Cross-Modal Analysis of Speech, Gestures, Gaze and Facial Expressions*. Springer, 2009, pp. 344–355.

[10] P. Birkholz, I. Steiner, and S. Breuer, "Control concepts for articulatory speech synthesis," in *the 6th ISCA Workshop on Speech Synthesis, Bonn, Germany, Proceedings*, 2007, pp. 5–10.

[11] J. Sun, X. Jing, and L. Deng, "Annotation and use of speech production corpus for building language-universal speech recognizers," in *Proc. of International Symposium on Chinese Spoken Language Processing*, vol. 3, 2000, pp. 31–34.

[12] H. Nam, L. Goldstein, E. Saltzman, and D. Byrd, "Tada: An enhanced, portable task dynamics model in matlab," *The Journal of the Acoustical Society of America*, vol. 115, no. 5, pp. 2430–2430, 2004.

[13] A. H. Wright, "Genetic algorithms for real parameter optimization," in *Foundations of genetic algorithms*. Elsevier, 1991, vol. 1, pp. 205–218.

[14] C. Z. Janikow and Z. Michalewicz, "An experimental comparison of binary and floating point representations in genetic algorithms." in *ICGA*, 1991, pp. 31–36.

[15] Y. Xu and F. Liu, "Tonal alignment, syllable structure and coarticulation: Toward an integrated model," *Italian Journal of Linguistics*, vol. 18, no. 1, p. 125, 2006.

[16] K. J. Kohler, "Zeitstrukturierung in der sprachsynthese," *ITG-Tagung Digitalc Sprachverarbeitung*, vol. 6, pp. 165–170, 1988.

[17] P. Birkholz, L. Martin, Y. Xu, S. Scherbaum, and C. Neuschaefer-Rube, "Manipulation of the prosodic features of vocal tract length, nasality and articulatory precision using articulatory synthesis," *Computer Speech & Language*, vol. 41, pp. 116–127, 2017.

[18] P. Birkholz, P. Schmaser, and Y. Xu, "Estimation of pitch targets from speech signals by joint regularized optimization," in *2018 26th European Signal Processing Conference (EUSIPCO)*. IEEE, 2018, pp. 2075–2079.

[19] S. J. Wright, "Coordinate descent algorithms," *Mathematical Programming*, vol. 151, no. 1, pp. 3–34, 2015.

[20] S. Radeck-Arneth, B. Milde, A. Lange, E. Gouvêa, S. Radomski, M. Mühlhäuser, and C. Biemann, "Open source german distant speech recognition: Corpus and acoustic model," in *International Conference on Text, Speech, and Dialogue*. Springer, 2015, pp. 480–488.

[21] T. Ellbogen, F. Schiel, and A. Steffen, "The bits speech synthesis corpus for german," *age*, vol. 47, no. 45, p. 40, 2004.

[22] K. W. DeReWo, "v-ww-bll-320000g-2012-12-31-1.0," 2013.

[23] B. Wu, Y. Xie, L. Lu, C. Cao, and J. Zhang, "The construction of a chinese interlanguage corpus," in *2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA)*. IEEE, 2016, pp. 183–187.

[24] T. Back, U. Hammel, and H.-P. Schwefel, "Evolutionary computation: Comments on the history and current state," *IEEE transactions on Evolutionary Computation*, vol. 1, no. 1, pp. 3–17, 1997.

[25] H.-G. Beyer and H.-P. Schwefel, "Evolution strategies – a comprehensive introduction," *Natural computing*, vol. 1, no. 1, pp. 3–52, 2002.

[26] P. Perrier and S. Fuchs, "11 motor equivalence in speech production," *The handbook of speech production*, p. 225, 2015.

[27] S. Dusan and L. Deng, "Acoustic-to-articulatory inversion using dynamical and phonological constraints," 2000, pp. 237–400.

[28] H. Kawahara, M. Morise, T. Takahashi, R. Nisimura, T. Irino, and H. Banno, "Tandem-straight: A temporally stable power spectral representation for periodic signals and applications to interference-free spectrum, f0, and aperiodicity estimation," in *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2008, pp. 3933–3936.

[29] J. Snoek, H. Larochelle, and R. P. Adams, "Practical bayesian optimization of machine learning algorithms," in *Advances in neural information processing systems*, 2012, pp. 2951–2959.