Araújo *et al. Journal of the Brazilian Computer Society* (2015) 21:17
DOI 10.1186/s13173-015-0037-9

Journal of
the Brazilian Computer Society
a SpringerOpen Journal

RESEARCH Open Access

Utterance copy through analysis-by-synthesis using genetic algorithm

Fabíola Araújo*, Aldebaro Klautau and José Filho

Abstract

Background: Utterance copy consists in estimating the input parameters to reconstruct a speech signal using a speech synthesizer. This process is distinct from the more traditional text-to-speech but yet used in many areas, especially in linguistics and health. Utterance copy is a difficult inverse problem because the mapping is non-linear and from many to one. It requires considerable amount of time to manually perform utterance copy and automatic methods, such as the one proposed here, are of interest.

Methods: This work presents our system based on genetic algorithm (GA) to automatically estimate the input parameters of the Klatt synthesizer using an analysis-by-synthesis process.

Results: Results are presented for synthetic (computer-generated) and natural (human-generated) speech, for male and female speakers. These results are compared with the ones obtained with WinSnoori, the only currently available software that performs the same task.

Conclusions: The experiments showed that the proposed *newGASpeech* system is an effective alternative to the laborious manual process of estimating the input parameters of a Klatt synthesizer. And it outperforms the baseline by a large margin with respect to five objective figures of merit. For example, in average, the mean squared error is reduced to approximately 60.4 % and 75.2 % when natural target voices from male and female speakers are used, respectively.

Keywords: Genetic algorithms; Utterance copy; Speech synthesis

Background

Utterance copy (or *copy synthesis*) corresponds to imitating the human voice via a synthesizer [1] and has important clinical applications [2]. For example, specialists use utterance copy to artificially produce the voices of patients who cannot normally speak due to trauma, disease, or surgery, and from the estimated set of parameters, they can better understand the related problems [2–4]. Thus, the task is that, given a target utterance (a sentence, word or phoneme spoken by the person of interest), one has to find the set of parameters that, when used as the input of a synthesizer, generates an artificial voice that resembles the target one. This task can be done manually, by trial-and-error, or automatically.

This paper presents a genetic algorithm called *newGASpeech* that performs the utterance copy task through a process of *analysis-by-synthesis* [2]. The obtained results are compared with the ones produced by the baseline WinSnoori [1], which to the best of the author's knowledge is the only freely available software that automatically estimates Klatt input parameters for utterance copy. Another related work is Procsy [5], but its current version requires additional input files, such as phonetic transcriptions that are not readily available.

The results presented in this work are for both *synthetic* speech, which is obtained with a text-to-speech (TTS) system, and *natural* speech, for male and female speakers. Due to the difficulty of an objective evaluation of the synthetic voices, several complementary figures of merit were adopted, namely: the log-spectral distance (D_{LE}), signal-to-noise ratio (SNR) [6], root-mean-square error (RMSE), Perceptual Evaluation of Speech Quality (PESQ) [7], and P.563, a single-ended method for objective speech quality [8].

*Correspondence: fpoliveira@ufpa.br
Signal Processing Laboratory of the Federal University of Pará, Rua Augusto Corrêa, - 01 – Guamá, Belém-PA, CEP 66075-110, Brazil

This work is organized as follows: background and a literature review about Klatt are described in the "Background" section. The methodology and customized genetic algorithm to perform utterance copy are explained in "Methods" and "Genetic algorithm for utterance copy" sections, respectively. The results and conclusions are presented in "Results and discussion" and "Conclusions" sections.

Klatt synthesizer

Klatt [9, 10] is a *formant*-based synthesizer adopted in many speech studies (e. g., [2, 11]) because most its inputs are closely related to physical parameters. This leads to a high degree of interpretability, which is essential in some studies of the acoustic correlates of voice quality, such as male/female speech conversion and simulation of breathiness, roughness, and vocal fry. The Klatt synthesizer has been used to mimic natural speech [12, 13] as well as pathological voices [2]. There are of course other synthesis techniques, together with methods for obtaining the associated input parameters, such as [14]. However, sometimes the interpretation of the role of each input parameter is not as easy as for Klatt and alternative synthesizers seem less popular.

The Klatt synthesizer uses a production mechanism based on a *source-filter* model [9, 15] that allows modeling the vocal tract through a linear filter, with a set of resonators in parallel or/and cascade that vary in time. There are several versions of Klatt synthesizers but the most popular ones are Klatt80 [9], *KLSYN88* [10] and a modified version of Klatt80 by Jon Iles [16]. Our system uses *KLSYN88* (Fig. 1), which has 48 parameters. All these parameters are detailed in [10] and, here, just a brief description is provided.

Only 41 *KLSYN88* parameters are effectively used here, and from the remaining seven, six of them are assumed to be zero (not shown in Fig. 1) and *SQ* is part of a voicing source that is not adopted in this work (Fig. 1, orange rectangle). The KLGLOTT88 voicing source was used and comprises *F*0 (fundamental frequency), *AV* (amplitude of voicing), *OQ* (open quotient), *FL* (flutter), and *DI* (diplophonia). The *TL* and *AH* parameters are parts of the voicing source and are responsible for an extra tilt of voicing spectrum and the amplitude of aspiration, respectively [10]. Five resonators in cascade are needed to simulate laryngeal sounds by the formants frequency F_1 to F_5 and their bandwidths B_1 to B_5, respectively. The resonators in parallel are responsible for modeling fricative sounds through parameters *A*2*F* to *A*6*F*, and for controlling the amplitude of fricative source *AF* and their bandwidths *B*2*F* to *B*6*F*.

The baseline WinSnoori is a free software for speech analysis that generates the Klatt parameters from the waveform input speech file. It makes utterance copy through a speech analysis algorithm. Its Klatt synthesizer is the modified version of Klatt80 by Jon Iles and having the configuration of 41 parameters to produce a speech frame.

The synthetic target speech files used in this study were generated by DECtalk [17], which is a *text-to-speech* (TTS) system produced by Fonix Corporation that internally uses *KLSYN88* as the backend synthesizer. The generated voices by DECtalk possess high intelligibility and its demo version was provided to LaPS (Signal Processing Laboratory of the Federal University of Pará) for research purposes. Because DECtalk generates speech using *KLSYN88*, it is very useful for our research given that, in this case, the "correct" Klatt input parameters are known, which does not occur when using natural speech.

More specifically, for each utterance specified by an input text, DECtalk generates an output file having 18 parameters that can be mapped to the 13 parameters of the *HLSyn* synthesizer [18] input file through a script developed in Java by our group and called *DEC2HLSyn*. The *HLSyn* can then be used to generate the input file of the *KLSYN88*, having the 48 necessary parameters to perform speech synthesis with Klatt. Generally speaking, *HLSyn* is a "high-level" synthesizer that expands its 13 input parameters into the 48 parameters of the "low-level" *KLSYN88* synthesizer [18].

Methods

In this study, 240 sentences [19] were submitted to the DECtalk TTS to produce synthetic voices for 6 different speakers (3 males and 3 females). Grouping the sentences into categories of male and female speakers, histograms of all parameters were generated and, from the histograms, it was possible to identify that DECtalk imposes variation of only 25 parameters (shown in black in Fig. 1), out of the 41 aforementioned, regardless of the speaker, while the others are kept at constant values. Parameters *FL*, *DF*1, *DB*1, and *A*6*F* have zero as constant value. The other constant parameters have their values listed in Table 1 and the parameters that vary over time are listed in Table 2. As in [10], only five resonators are adopted in this work, such that *F*6 and *B*6 are not used. The suggested range of parameter values defined in [10] is expanded by DECtalk for parameters *AF*, *B*1, *FNP*, and *FNZ* [20].

The KLSYN88 input file is formed by several rows, each one representing a voice segment and having the combination of 25 different parameter values, which are used to synthesize the synthetic speech for that frame in this work (Fig. 2).

The production of a KLSYN88 input file needs considerable time if performed manually. As discussed, the goal of *newGASpeech* is, given a waveform file with the target spoken utterance, to automatically estimate the input parameters of Klatt synthesizer. It uses a GA based on the

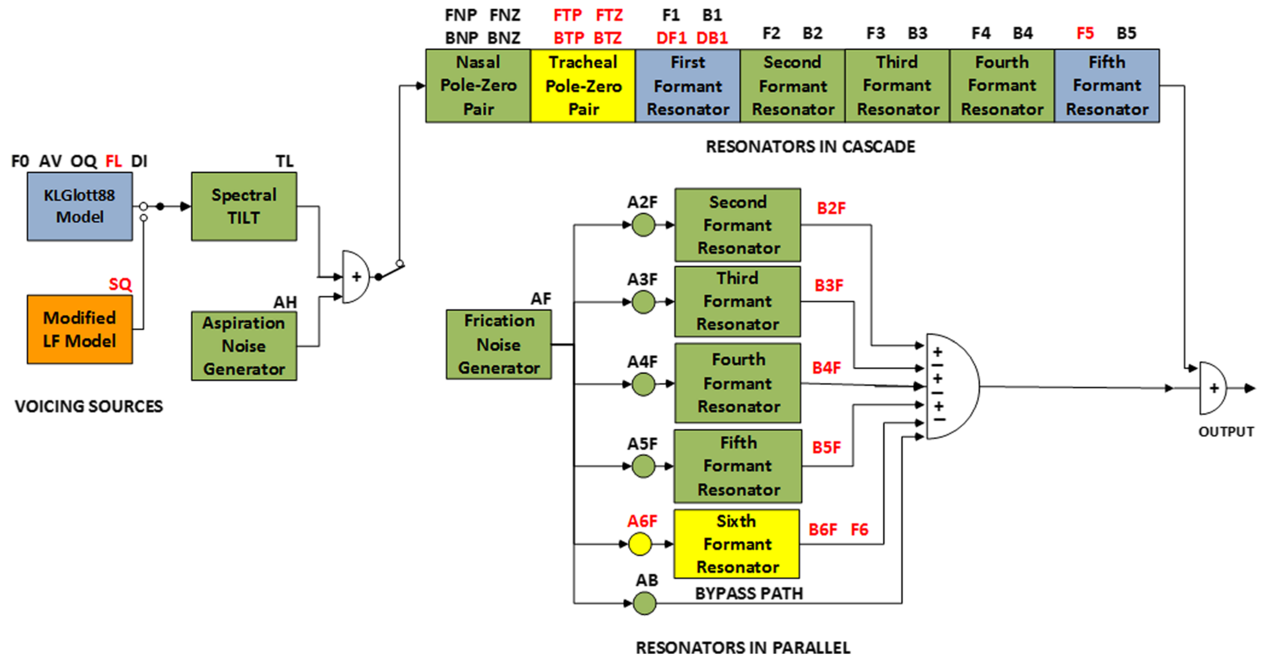


Fig. 1 The KLSYN88 formant synthesizer adapted from [10]

Non-Dominated Sorting Genetic Algorithm II (NSGA-II) [21] with the Root Mean Squared Error (RMSE) as fitness function to evaluate the solution. The time to execute *newGASpeech* on a long utterance is significant, but the process is feasible because it can be done offline and does not demand time from a specialist to perform a manual utterance copy.

Genetic algorithm for utterance copy

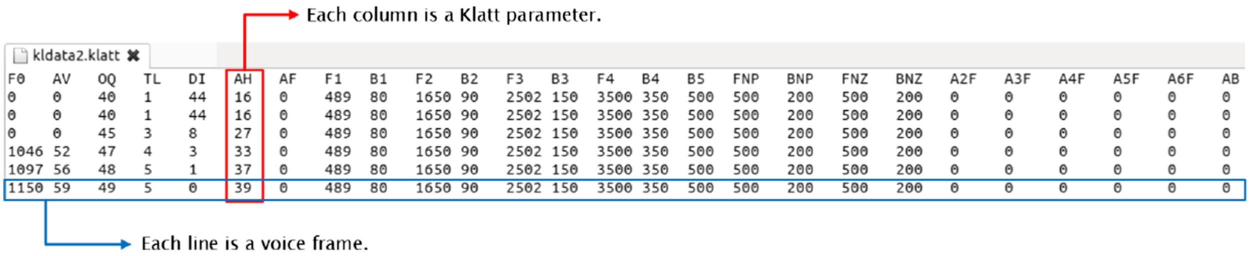
The analysis-by-synthesis process begins by segmenting the input target voice file into frames of approximately 5 ms and obtaining information through a configuration file as shown in Fig. 3 (step 1). For each frame, a full run of GA is executed (several iterations). The parameters to synthesize the frame compose a chromosome that has its fitness calculated after using its corresponding parameters to feed Klatt. The fitness is the RMSE calculated between the target and synthesized signal frames. Care has to be exercised to properly re-initialize the synthesizer's state (memory of its digital resonators, etc.) along the whole process that executes several GAs. After evaluating the population, a rank is assigned to each individual and those with better ranks are selected to undergo crossover and mutation. As a result, a new population is generated and, in turn, undergoes evaluation, selection, crossover, and mutation steps again. The whole process is repeated until the algorithm reaches one of the stopping criteria

Table 1 Klatt parameters with constant values different from zero

Parameter	Value	Parameter	Value
F5	4500	B2F	250
FTP	1000	B3F	320
BTP	200	B4F	350
FTZ	1000	B5F	500
BTZ	200	B6F	1500

Table 2 25 *KLSYN88*'s changing parameters

P.	Min	Max	Unit	P.	Min	Max	Unit
F0	0	5000	Hz	F4	2400	4990	Hz
AV	0	80	dB	B4	3000	4990	Hz
OQ	0	99	%	B5	100	1500	Hz
TL	0	41	dB	FNP	450	870	Hz
DI	0	100	%	BNP	40	1000	Hz
AH	0	80	dB	FNZ	180	1000	Hz
AF	0	70	dB	BNZ	40	1000	Hz
F1	180	1300	Hz	A2F	0	80	dB
B1	30	1040	Hz	A3F	0	80	dB
F2	550	3000	Hz	A4F	0	80	dB
B2	40	1000	Hz	A5F	0	80	dB
F3	1200	4800	Hz	AB	0	80	dB
B3	60	1000	Hz	–	–	–	–


Each column is a Klatt parameter.
kldata2.klatt
Each line is a voice frame.

```
F0   AV   OQ   TL   DI   AH   AF   F1   B1   F2   B2   F3   B3   F4   B4   B5   FNP  BNP  FNZ  BNZ  A2F  A3F  A4F  A5F  A6F  AB
0    0    40   1    44   16   0    489  80   1650 90   2502 150  3500 350  500  500  200  500  200  0    0    0    0    0    0
0    0    40   1    44   16   0    489  80   1650 90   2502 150  3500 350  500  500  200  500  200  0    0    0    0    0    0
0    0    45   3    8    27   0    489  80   1650 90   2502 150  3500 350  500  500  200  500  200  0    0    0    0    0    0
1046 52   47   4    3    33   0    489  80   1650 90   2502 150  3500 350  500  500  200  500  200  0    0    0    0    0    0
1097 56   48   5    1    37   0    489  80   1650 90   2502 150  3500 350  500  500  200  500  200  0    0    0    0    0    0
1150 59   49   5    0    39   0    489  80   1650 90   2502 150  3500 350  500  500  200  500  200  0    0    0    0    0    0
```

Fig. 2 KLSYN88 input file

[22] (step 2). The best individuals for each frame compose a Klatt input file which is synthesized and outputs a synthetic voice file that aims at mimicking the target voice (steps 3 to 6).

In most cases, the speech signal does not change abruptly from one frame to another. In order to improve convergence and suggest continuity with respect to the previous population, the framework has the option of running *Interframe* [22]. In this case, the best individuals of the previous frame are copied to initialize the population of the next frame (recall that a new GA-population is initialized for each frame).

Coding and decoding of chromosome

The population is comprised of individuals with chromosomes that have genes which represent the Klatt synthesizer input parameters of a frame. Genes are grouped in two sections: F_v and T_v that store the parameters of voicing source and vocal tract, respectively. Figures 4 and 5 illustrate all genes that may comprise F_v and T_v sessions, respectively.

In addition to the sections mentioned above, the chromosome also has a voicing gene. This gene performs the necessary function of indicating if the chromosome represents a voiced or unvoiced segment. If the segment

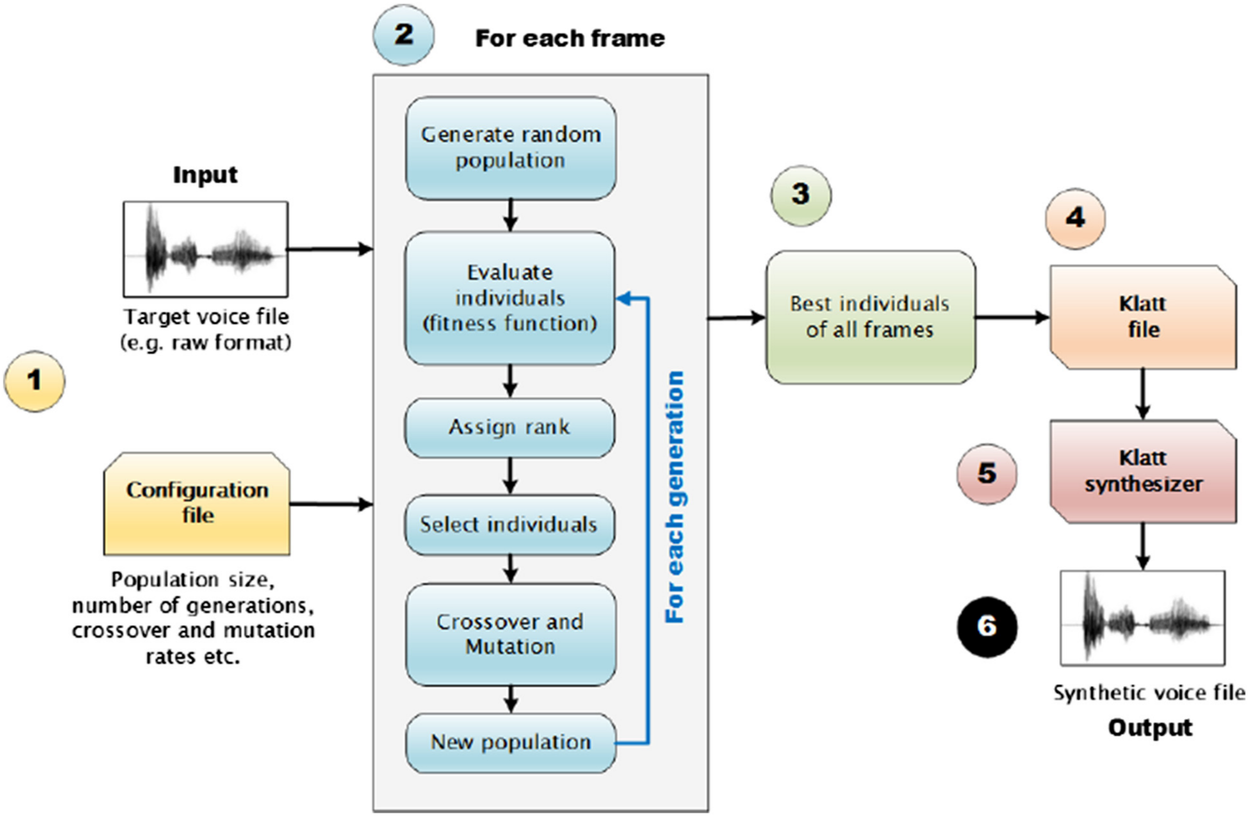

2
For each frame
Generate random population
Evaluate individuals (fitness function)
Assign rank
Select individuals
Crossover and Mutation
New population
For each generation
Input
Target voice file (e.g. raw format)
1
Configuration file
Population size, number of generations, crossover and mutation rates etc.
3
Best individuals of all frames
4
Klatt file
5
Klatt synthesizer
6
Synthetic voice file
Output

Fig. 3 Description of the *newGASpeech* framework

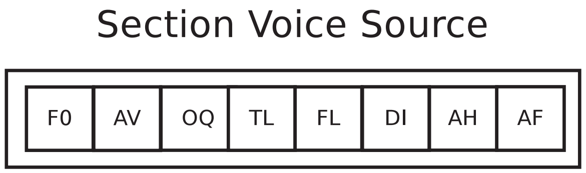


Fig. 4 Structure of the voicing source section

is unvoiced, $F0$ and AV parameters are zero, otherwise the sound is identified as voiced and both have nonzero values.

Look-ahead mechanism

Usually, the speech signal synthesized by Klatt is composed of several frames. A frame should not be treated independently because it might potentially influence the next frames. Therefore, a frame configuration may be good for the current frame; although, it may impact negatively on the signal of the following frames. The Look-Ahead mechanism [20] allows the evaluation of the synthesis of the current frame configuration together with n_f following frames, greatly increasing the problem search space. Figure 6 illustrates the chromosome structure.

B_v is the voicing source gene, F_v, T_v, and L_a are the sections: voicing source, the vocal tract, and Look-Ahead mechanism, respectively. B_v, F_v, and T_v were previously addressed. The L_a section stores Look-Ahead frames and its size depends on the amount of k frames needed ahead to solve the problem. Experiments demonstrated in this study that to estimate the 25 parameters mentioned in the "Methods" section at least $n_f = 2$ Look-Ahead frames are needed with all its frames having the same weight equal to one. Thus, L_a section contains at least B_v, F_v, and T_v for the next 2 frames.

This minimum amount of Look-ahead frames is based on the time interval t_0 between impulses to generate a voiced excitation. Hence, the number of samples T_0 corresponding to t_0 seconds is $T_0 = t_0 f_s$, where f_s is the sample rate in Hz. For any periodic signal, the fundamental frequency $f_0 = 1/t_0$ (in Hz) is one over the fundamental period t_0. Hence, to obtain T_0 as a function of the parameter $F0$, it remains to note that Klatt uses an integer parameter $F0 = 10f_0$ to represent f_0 [10]. Therefore, $T_0 = f_s/(F0/10)$ is the approximate number of samples separating voicing impulses. In this research, $f_s = 11025$ and each frame is represented by 71 samples. Figure 7 shows the signal of the voicing source to 4 frames where N is the current. In this case, the frame N has $F0$ value equal to 943 and, according to the T_0 equation, the next impulse will occur 116.9 samples ahead from the beginning of the period in this frame, in other words in frame $N + 2$.

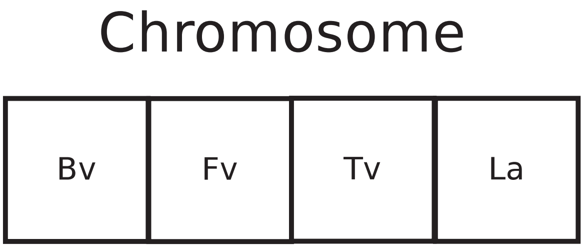


Fig. 6 Division of chromosome into sections

The $F0$ average value in the experiments was 975.5 for male voiced frames, therefore, the $F0$ value set for the current frame only impacts in the second follows and the frame N is as important as other. For female speakers, the $F0$ average value was greater ($\approx$ 1595.5) than male. Applying the T_0 equation to calculate the number of samples for the next impulse were required approximately 69.1 samples. This indicates that for female speakers only $N + 1$ Look-ahead frame is required because the $F0$ value impacts directly in the next frame.

Dimensionality of the search space

Each Klatt parameter has its own possible range of values [10], with these values restricted to be integer numbers. For a single frame, the size of the search space S is given by:

$$S = \prod_{n=1}^{N_p} (U_n - L_n + 1) \tag{1}$$

where N_p is the number of parameters to be estimated and U_n and L_n, respectively, represent the upper and lower bounds of integer and consecutive values of parameter

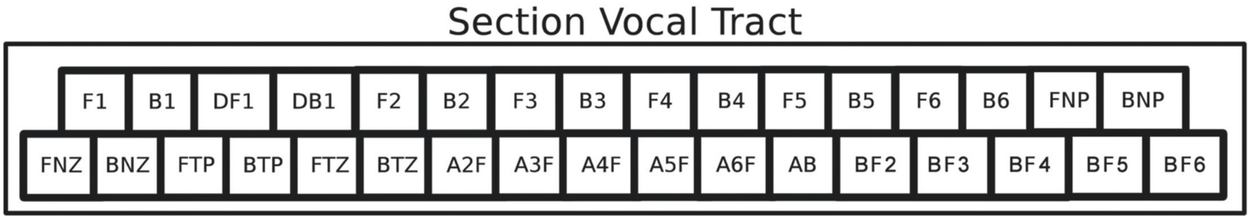


Fig. 5 Structure of the vocal tract section of a chromosome

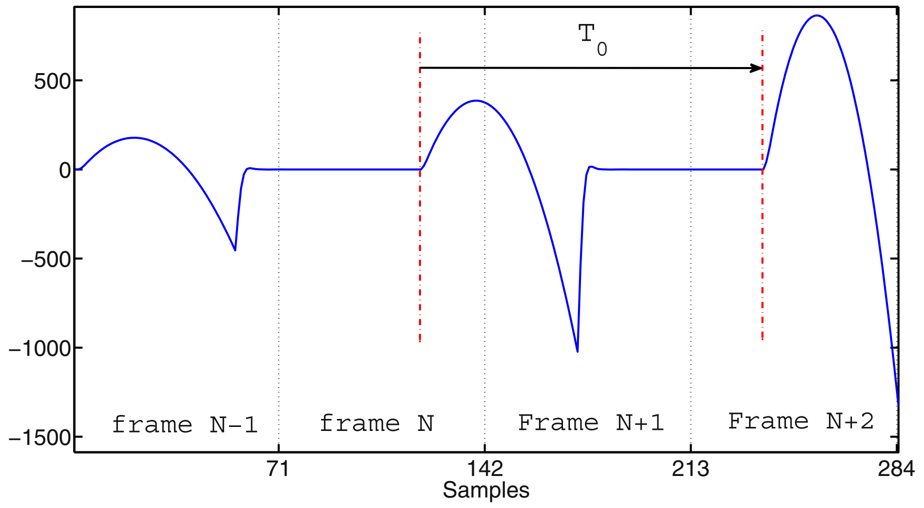


Fig. 7 Impulses of a voiced excitation

n. For example, if the framework is estimating $N_p = 25$ parameters for one frame, and each can assume $U_n - L_n + 1 = 50$ distinct values, $\forall n$, the search space is given by $S = (51)^{25} \approx 5 \times 10^{42}$. This dimension gets even larger when the search space includes the Look-Ahead frames and is then given by $S^{(n_f+1)}$. Hence, procedures such as GA are an important tool to address this problem.

Results and discussion

Utterance copy recent experiments were performed in two ways: first to assess the GA convergence regarding the dimensionality of the search space and second to compare the synthesized male and female voices obtained by the *newGASpeech* and the WinSnoori using as target natural and synthetic voices. The main motivation to use synthetic and natural voices was to have finer control of the experiments in the former, given that the correct values of the input parameters are known, and test our system with natural voices of unknown speakers in the latter. The DECtalk voices were from six American male and female speakers and the words were listed in the Table 3.

The used natural voices were from the TIDIGITS corpus [23]. These voices are pronunciations by six American male and female speakers of the digits: *two*, *four*, *six*, *eight*, and *nine* totaling 30 speech signals.

Table 3 Male and female words list—synthetic voices

Index	Frank	Harry	Paul	Betty	Ursula	Wendy
1	air	awe	are	air	bean	death
2	dill	earl	end	dill	earl	fern
3	hurl	gill	is	hurl	tang	is
4	jam	them	no	jam	them	then
5	who	wish	there	who	wish	there

The *newGASpeech* was configured as shown in Table 4. The crossover and mutation rates are adaptive and may be decremented in 0.01 by each generation. This occurs until the minimum rate equal to 0.01 is reached. However, the rates are only decreased if the population presents *diversity*. The option of running *Interframe* was used and 10 % of the best chromosomes from the previous frame were copied to initialize part of the next frame population. For the following simulations, GA was configured to always estimate the 25 varying parameters (“Methods” section).

Files generated by DECtalk and from TIDIGITS corpus were used as input to both *newGASpeech* and WinSnoori. The target and synthetic signals were then aligned according to their cross-correlation and the results evaluated using the metrics: SNR, RMSE, D_{LE}, PESQ, and P.563. It should be noted that none of these metrics perfectly correlate with subjective evaluations. However, informal listening tests indicated a very good match between the overall result of the four metrics and a MOS-like subjective evaluation.

The SNR value should be as large as possible indicating that the signal power is much higher than the “noise” power. While for RMSE and D_{LE}, the smaller the better. PESQ and P.563 compares two speech signals and assigns scores ranging from 1 to 5, with 1 being “bad quality” and 5 “excellent quality”. An important difference between these

Table 4 *newGASpeech* configuration

Parameter	Value
Number of generations	5000
Population size	800
Initial crossover rate	50 %
Initial mutation rate	30 %

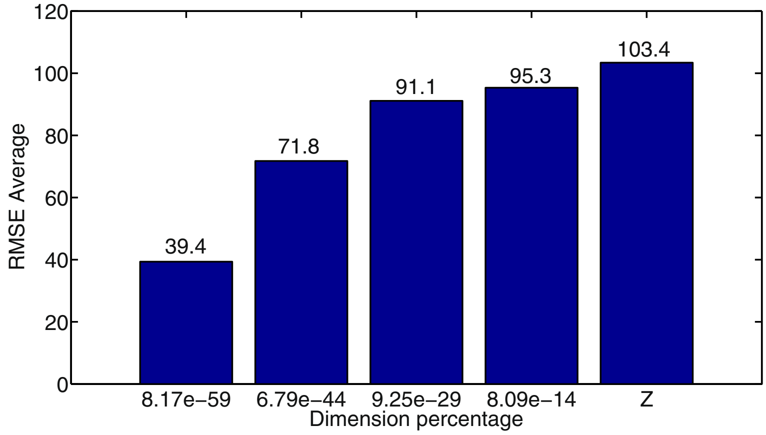


Fig. 8 RMSE average for five words of Paul speaker

last two is that P.563 is single-ended, i.e., it does not use the target file when assigning a score. This is interesting in utterance copy because the other four metrics require a reasonable time-alignment between the target and synthetic signals while P.563 does not. However, the P.563 is not appropriate for speech signals with duration shorter than 3 s [8] and, because of that, it was not used to evaluate the signals from TIDIGITS.

Dimensionality of the search space

To evaluate how dimensionality influences the fitness function (RMSE), the option of informing was developed frame by frame and, for each parameter, a restricted range of values around the correct one. Experiments were performed using five words uttered by one of the male speakers shown in Table 3, and the *newGASpeech*'s configuration was as follows: population of 1000 individuals, 300 generations, and the same initial crossover and mutation rates specified in the Table 4. From the correct parameters values for a given frame, they were varied in ±2, 4, 8, 16, and 32 %.

In this experiment, the largest search space dimension occurs when 25 parameters are estimated and the correct parameter values are restricted to the variation of ±32 %. The variable Z is this dimension and was calculated by Eq. 1. Its value is approximately 6.86E + 102, including the dimension of the 2 Look-ahead frames required. Other dimensions of the search space, for ±2, 4, 8, and 16 % variations of the parameter values, were normalized by Z. Figure 8 illustrates the average RMSE calculated for all words. It can be seen that the search space is huge and the RMSE only decreases significantly when it is reduced by several orders of magnitude.

Experiments with synthetic target voices

Figures 9 and 10 show the results for synthetic speech using "boxplot" graphs, for male and female speakers. In all performed tests, the result obtained by *newGASpeech* was better for all speakers than the one by WinSnoori according to the chosen figures of merit, with those of *Wendy* with larger difference in favor of *newGASpeech* with lower RMSE (91.6 %), greater PESQ (28.4 %), SNR (99.6 %), and P.563 (78.3 %) medians although the D_{LE} has not been the lowest (91.2 %). These percentages reflect the variation with respect to the baseline. All female and male speakers obtained similar results when

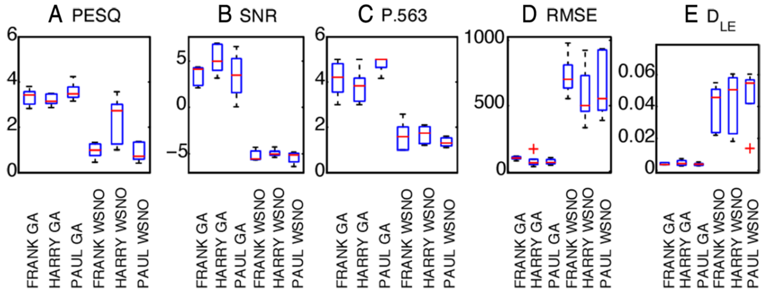


Fig. 9 **a** PESQ, **b** SNR, **c** P.563, **d** RMSE, and **e** D_{LE} values for male speakers

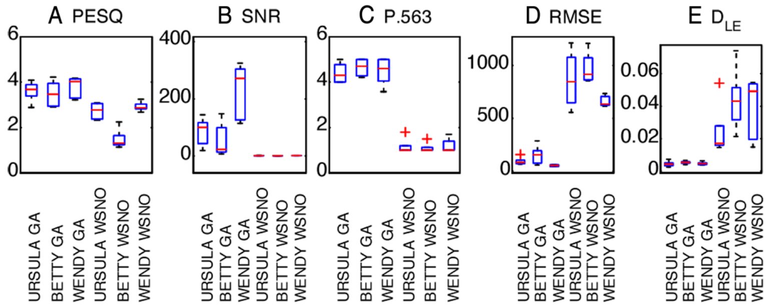


Fig. 10 **a** PESQ, **b** SNR, **c** P.563, **d** RMSE, and **e** D_{LE} values for female speakers

using GA (Table 5), except for the SNR average which was 96.8 % higher for female than for male.

Experiments with natural target voices

In experiments using natural target voices, as seen in Figs. 11 and 12, the GA obtained voices with higher PESQ and SNR, and lower RMSE and D_{LE} than the baseline. Results of female speakers performed better than male. The female *Speaker 2* presents the best PESQ median value (3.19) although the RMSE (40.21) and D_{LE} (0.028) of the *newGASpeech* was not the lowest value and the SNR (21.45) not be the highest if compared to other speakers.

Comparing the results obtained with synthetic and natural target voices (Table 5), the *newGASpeech* had a greater RMSE reduction of 43.5 % and 35.7 % for female and male speakers although the PESQ reduced too. The SNR increased for male and decreased for female voices, and the D_{LE} increased significantly for both ($\approx$ 88 % in average). The P.563 value is considered high for the results with synthetic target voices with an average value of 4.4 for male and female, but of course the results with natural speech are more important.

For all experiments, WinSnoori was outperformed. Comparing the results of the baseline for synthetic versus natural target voices, the PESQ median reduced $\approx$14 % for female and increased the same percentage for male speakers. The SNR increased a little, and as in the GA, the major difference was the reduction of the RMSE, $\approx$75 % average, for male and female, and increase of the D_{LE} in $\approx$86 % for both.

Percentage error of the *newGASpeech* estimated parameters for the synthetic voices

Given that the correct values were known, for all experiments with male and female synthetic voices (Table 3) using GA, the percentage error (*PE*) was calculated as

$$PE = 100 \times \left|\frac{\hat{p_k} - p_k}{p_k}\right| \% \tag{2}$$

for all 25 estimated parameters, where $\hat{p_k}$ is the estimated Klatt parameter value by GA and p_k its target value. The goal is to observe, in the input parameters space, what are the parameters with the best and worst estimations.

For each word, the *PE* average was calculated for voiced frames to avoid silence and low-energy unvoiced frames. Female and male speakers had seven parameters with an average *PE* less than 10 % and this same number of parameters with average error in the range between 10 % and 50 %. These parameters belong to the voicing source and cascade branch and are listed in Table 6.

The parameters with the smallest PE in Table 6 are those that impact the synthesized speech the most. For example, the parameters belonging to the parallel branch, responsible for the production of fricative sounds, had $PE >$

Table 5 Medians for *newGASpeech* (GA) and WinSnoori (Wsno) by voice type for male speakers

	Male				Female			
	Natural voice		Synthesized voice		Natural voice		Synthesized voice	
Metrics/median	GA	Wsno	GA	Wsno	GA	Wsno	GA	Wsno
RMSE	50.3	126.9	78.3	585.6	54.8	221.2	97	799
D_{LE}	0.04	0.38	0.004	0.05	0.028	0.26	0.004	0.036
PESQ	2.5	1.7	3.4	1.5	2.9	2.0	3.7	2.3
P.563	–	–	4.4	1.5	–	–	4.5	1
SNR	8.6	0.3	4.2	−5.2	15.1	1.2	131.4	0.9

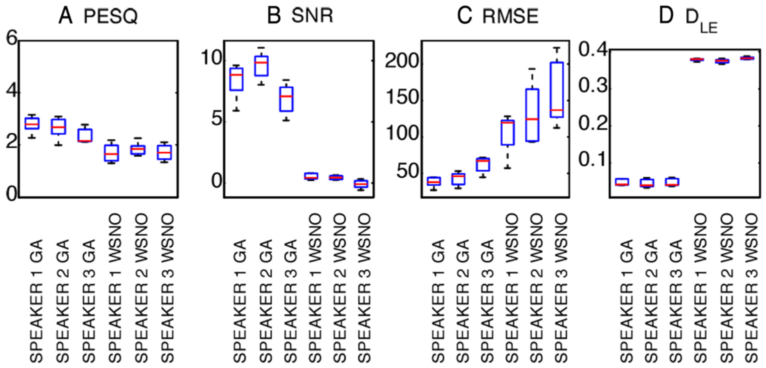


Fig. 11 **a** PESQ, **b** SNR, **c** RMSE, and **d** D_{LE} values for male speakers

100 % indicating that these parameters did not strongly influence in the distortion of the generated signal. In contrast, $F1$ was the parameter with highest accuracy for male speakers (98 %) given its strong impact on the generated speech. Curiously, for all speakers, AH and TL were in the same PE range.

Conclusions

This work presented the current version of the *newGASpeech* framework, which is based on GA and performs utterance copy through a process of *analysis-by-synthesis*. The obtained results were compared with the ones produced by the baseline WinSnoori.

The proposed software significantly outperformed WinSnoori with respect to five objective figures of merit: RMSE, SNR, spectral distortion, PESQ, and P.563 scores. The results were obtained for synthetic and natural speech files covering both male and female speakers. Compared to previous results, the RMSE decreased by 72.8 % and 64.8 % for male and female speakers, respectively.

This is on-going work with improvements to be made. For example, after the systems are properly tuned, a systematic subjective evaluation will be conducted. Another aspect that was taken in account in the design of the current experiments is that breadth, instead of depth, was prioritized. Future experiments will adopt larger amount of speech data, especially with relatively long sentences.

Another important future work is to evaluate the results according to the input parameters themselves (their time evolution, dynamic range, etc.). Because the problem has a many-to-one mapping, it happens that a solution has a good match with the target speech when only the synthetic speech is taken into account, but a relatively poor set of input parameters.

The *newGASpeech* will be made freely available for users of Klatt-based utterance copy applications. The final goal is to provide an easy-to-use solution that focuses on utterance copy for health applications. In spite of having more than three decades, the Klatt synthesizer is still a popular formant synthesizer and this motivates the development of associated tools.

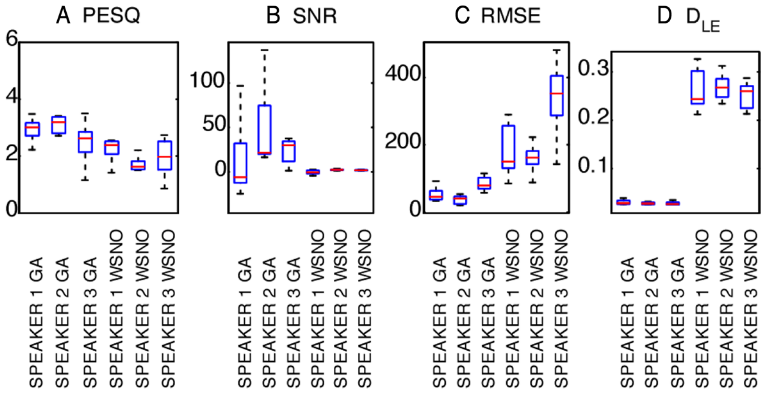


Fig. 12 **a** PESQ, **b** SNR, **c** RMSE, and **d** D_{LE} values for female speakers

Table 6 Parameters with percentage error below 50 %

Branch	Parameters		
	PE	Male	Female
Voicing source	$PE < 10\,\%$	AV and OQ	F0, AV, and OQ
	$10\,\% < PE < 50\,\%$	AH and TL	AH and TL
Cascade branch	$PE < 10\,\%$	F1, F2, F4, B4, and B5	F4, FNP, B4, and B5
	$10\,\% < PE < 50\,\%$	F3, B3, FNP, BNP, and FNZ	F1, F2, F3, B3, and BNP

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

FA is the main author of this paper, which is the result of her doctoral thesis in the Electrical Engineering Graduate Program at Federal University of Pará. Besides her academic research that comprises the literature review, she led the development team, made simulations and collected all data results of the genetic algorithm and baseline software. AK is FA's research advisor. He was responsible for proposing the problem and guides the research. JF is a Master's student who participated in the results discussions and is part of the research group. All authors have contributed in drafting, writing and revising the manuscript and agree to be accountable for all aspects of the work, thus taking public responsibility for the content. All authors read and approved the final manuscript.

Authors' information

FA obtained her Master's degree in Computer Science at the Federal University of Pernambuco (Recife-PE, Brazil) and is professor at the Federal University of Pará (Castanhal-PA, Brazil). AK obtained his Ph.D. degree in Electrical Enginnering at University of California (San Diego, USA) and is a professor at the Federal University of Pará (Belém-PA, Brazil). JF is a Master's student in Computer Science at the Federal University of Pará (Belém-PA, Brazil) and graduated in Computer Science at Centro Universitário do Estado do Pará.

Author details

Received: 5 September 2014 Accepted: 21 September 2015
Published online: 08 October 2015

References

1. Laprie Y (2002) The WinSnoori user's manual version 1.32. Manuel technique. hal.inria.fr/inria-00107630
2. Bangayan P, Long C, Alwan AA, Kreiman J, Gerratt BR (1997) Analysis by synthesis of pathological voices using the Klatt synthesizer. In: Speech Communication, vol 22. Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands. pp 343–368
3. Kain E, Niu X, Hosom JP, Miao Q, Santen JV (2004) Formant resynthesis of dysarthric speech. In: IEEE Workshop on Speech Synthesis. pp 25–30
4. Fraj S, Schoentgen J, Grenez F (2012) Development and perceptual assessment of a synthesizer of disordered voices. In: The Journal of the Acoustical Society of America, vol 132. Acoustical Society of America, Suite 300 1305 Walt Whitman Road Melville, NY 11747-4300. pp 2603–2615
5. Heid S, Hawkins S (1998) PROCSY: A hybrid approach to high-quality formant synthesis using HLsyn. In: Third International Workshop on Speech Synthesis. Jenolan Caves, Australia. pp 219–224
6. Al-Akhras M, Daqrouq K, Al-Qawasmim AR (2010) Perceptual evaluation of speech enhancement. In: Systems Signals and Devices (SSD), 2010 7th International Multi-Conference. Institute of Electrical and Electronics Engineers. pp 1–6
7. ITU-T Recommendation P.862 (2001) Perceptual Evaluation of Speech Quality (PESQ): An Objective Method for End-To-End Speech Quality Assessment of Narrow-Band Telephone Networks and Speech Codecs. www.itu.int/rec/T-REC-P.862. Accessed 20 January 2014
8. ITU-T Recommendation P.563 (2004) Single-ended method for objective speech quality assessment in narrow-band telephony applications. www.itu.int/rec/T-REC-P.563-200405-I/en. Accessed 02 July 2014
9. Klatt D (1980) Software for a cascade/parallel formant synthesizer, Vol. 67. Acoustical Society of America, Suite 300 1305 Walt Whitman Road Melville, NY 11747-4300
10. Klatt D, Klatt L (1990) Analysis, synthesis, and perception of voice quality variations among female and male speakers, Vol. 87. Acoustical Society of America, Suite 300 1305 Walt Whitman Road Melville, NY 11747-4300
11. Jinachitra P, Smith III JO (2005) Joint estimation of glottal source and vocal tract for vocal synthesis using Kalman smoothing and EM algorithm. In: Applications of Signal Processing to Audio and Acoustics. Institute of Electrical and Electronics Engineers. pp 327–330
12. Anumanchipalli GK, Cheng YC, Fernandez J, Huang X, Mao Q, Black AW (2010) KlaTTStat: Knowledge-based Statistical Parametric Speech Synthesis. In: 7th ISCA Workshop on Speech Synthesis
13. Shrivastav R, Sapienza CM (2006) Some difference limens for the perception of breathiness, Vol. 120. Acoustical Society of America, Suite 300 1305 Walt Whitman Road Melville, NY 11747-4300
14. Liu C, Kewley-Port D (2004) STRAIGHT: A new speech synthesizer for vowel formant discrimination. In: Acoustic Research Letters Online. pp 31–36
15. Lemmetty S (1999) Review of Speech Synthesis Technology. Master's thesis, Department Electrical and Communication Engineering - Helsinki University of Technology, Finland
16. Laprie Y, Bonneau A (2002) A copy synthesis method to pilot the Klatt synthesizer. In: 7th International Conference on Spoken Language Processing. ICSLP2002 - INTERSPEECH 2002, Denver, Colorado, USA. p 4
17. Hallahan WY (1995) DECtalk Software: Text-to-Speech Technology and Implementation. In: Digital Technical Journal, vol 7. pp 5–19
18. Hanson HM, McGowan RS, Stevens KN, Beaudoin RE (1999) Development of Rules for Controlling the HLSyn Speech Synthesizer. In: Proceedings of the Acoustics, Speech, and Signal Processing, vol 1. Institute of Electrical and Electronics Engineers. pp 85–88
19. Rothauser EH, Chapman WD, Guttman N, Hecker MHL, Nordby KS, Silbiger HR, Urbanek GE, Weinstock M (1969) IEEE recommended practice for speech quality measurements. In: IEEE Transactions on Audio and Electroacoustics, vol 17. Institute of Electrical and Electronics Engineers. pp 225–246
20. Trindade J, Araujo F, Klautau A, Batista P (2013) A genetic algorithm with look-ahead mechanism to estimate formant synthesizer input parameters. In: IEEE Congress on Evolutionary Computation. Institute of Electrical and Electronics Engineers. pp 3035–3042
21. Srinivas N, Deb K (1994) Multiobjective Optimization Using Nondominated Sorting in Genetic Algorithms. In: Evolutionary Computation, vol 2. Massachusetts Institute of Technology, MIT Press Cambridge, MA, USA. pp 221–248
22. Oliveira F, Trindade J, Borges J, Couto I, Klautau A (2011) Multi-objective genetic algorithm to automatically estimating the input parameters of formant-based speech synthesizers. In: INTECH Open Access Publisher, vol 2. INTECH Open Access Publisher. pp 283–302
23. Leonard RG, Doddington G (1993) TIDIGITS. catalog.ldc.upenn.edu/LDC93S10. Accessed 10 March 2014