# Interactive Intonation Optimisation Using CMA-ES and DCT Parameterisation of the F0 Contour for Speech Synthesis

Adriana STAN, Florin-Claudiu POP, Marcel CREMENE,
Mircea GIURGIU, Denis PALLEZ

**Abstract** Expressive speech is one of the latest concerns of text-to-speech systems. Due to the subjectivity of expression and emotion realisation in speech, humans cannot objectively determine if one system is more expressive than the other. Most of the text-to-speech systems have a rather flat intonation and do not provide the option of changing the output speech. We therefore present an interactive intonation optimisation method based on the pitch contour parameterisation and evolution strategies. The Discrete Cosine Transform (DCT) is applied to the phrase level pitch contour. Then, the genome is encoded as a vector that contains 7 most significant DCT coefficients. Based on this initial individual, new speech samples are obtained using an interactive Covariance Matrix Adaptation Evolution Strategy (CMA-ES) algorithm. We evaluate a series of parameters involved in the process, such as the initial standard deviation, population size, the dynamic expansion of the pitch over the generations and the naturalness and expressivity of the resulted individuals. The results have been evaluated on a Romanian parametric-based speech synthesiser and provide the guidelines for the setup of an interactive optimisation system, in which

Adriana STAN
Communications Department, Technical University of Cluj-Napoca, Cluj, Romania,
e-mail: Adriana.Stan@com.utcluj.ro

Florin-Claudiu POP
Communications Department, Technical University of Cluj-Napoca, Cluj, Romania,
e-mail: Florin.Pop@com.utcluj.ro

Marcel CREMENE
Communications Department, Technical University of Cluj-Napoca, Cluj, Romania,
e-mail: Cremene@com.utcluj.ro

Mircea GIURGIU
Communications Department, Technical University of Cluj-Napoca, Cluj, Romania,
e-mail: Mircea.Giurgiu@com.utcluj.ro

Denis PALLEZ
Laboratoire d'Informatique, Signaux, et Systèmes de Sophia-Antipolis (I3S), Université de Nice Sophia-Antipolis, France, e-mail: Denis.Pallez@unice.fr

the users can subjectively select the individual which best suits their expectations with minimum amount of fatigue.

## 1 Introduction

Over the last decade text-to-speech (TTS) systems have evolved to a point where in certain scenarios, non-expert listeners cannot distinguish between human and synthetic voices with 100% accuracy. One problem still arises when trying to obtain a natural, more expressive sounding voice. Several methods have been applied ([17], [20]), some of which have had more success than others and all of which include intonation modelling as one of the key aspects. Intonation modelling refers to the manipulation of the pitch or fundamental frequency (F0). The expressivity of speech is usually attributed to a dynamic range of pitch values. But in the design of any speech synthesis system (both concatenative and parameteric), one important requirement is the flat intonation of the speech corpus, leaving limited options for the synthesised pitch contours.

In this paper we propose an interactive intonation optimisation method based on evolution strategies. Given the output of a synthesiser, the user can opt for a further enhancement of its intonation. The system then evaluates the initial pitch contour and outputs a small number of different versions of the same utterance. Provided the user subjectively selects the best individual in each set, the next generation is built starting from this selection. The *dialogue* stops when the user considers one of a generation's individual satisfactory. The solution for the pitch parameterisation is the Discrete Cosine Transform (DCT) and for the interactive step, the Covariance Matrix Adaptation-Evolution Strategy (CMA-ES).

This method is useful in the situation where non-expert users would like to change the output of a speech synthesiser to their preference. Also, under resourced languages or limited availability of speech corpora could benefit from such a method. The prosodic enhancements selected by the user could provide long-term feedback for the developer or could lead to a *user-adaptive* speech synthesis system.

### *1.1 Problem statement*

In this subsection we emphasise some aspects of the current state-of-the-art speech synthesisers which limit the expressiveness of the result:

**Issue #1**: Some of the best TTS systems benefit from the prior acquisition of a large speech corpus and in some cases extensive hand labelling and rule-based intonation. But this implies a large amount of effort and resources, which are not available for the majority of languages.

**Issue #2**: Most of the current TTS systems provide the user with a single unchangeable result which can sometimes lack the emphasis or expressivity the user might have hoped for.

**Issue #3:** If the results of a system can be improved, it usually implies either additional annotation of the text or a trained specialist required to rebuild most or all of the synthesis system.

**Issue #4:** Lately, there have been studies concerning more objective evaluations of the speech synthesis, but in the end the human is the one to evaluate the result and this is done in a purely subjective manner.

### *1.2 Related work*

To the best of our knowledge, evolution strategies have not been previously applied to speech synthesis. However, the related genetic algorithms have been used in articulatory [1] or neural networks based [11] speech synthesisers. A study of interactive genetic algorithms applied to emotional speech synthesis is presented in [8]. The authors use the XML annotation of prosody in Microsoft Speech SDK and try to convert neutral speech to one of the six basic emotions: *happiness, anger, fear, disgust, surprise* and *sadness*. The XML tags of the synthesised speech comprise the genome. Listeners are asked to select among 10 speech samples at each generation and to stop when they consider the emotion in one of the speech samples consistent with the desired one. The results are then compared with an expert emotional speech synthesis system. Interactive evolutionary computation has, on the other hand, been applied to music synthesis [10],and music composition [3], [9].

## 2 DCT parameterisation of the F0 Contour

In text-to-speech one of the greatest challenges remains the intonation modelling. There are many methods proposed in order to solve this problem, some taking into account a phonological model [2], [15] and others simply parameterising the pitch as a curve [18]. Curve parameterisation is a more efficient method in the sense that no manual annotation of the text to be synthesised is needed and thus not prone to subjectivity errors.

Because in this study we are not using prior text annotations or additional information, we chose a parameterisation based on the DCT, that partially adresses Issue #1 of the Problem Statement section.

DCT is a discrete transform which expresses a sequence of discrete points as a sum of cosine functions oscillating at different frequencies with zero phase. The are several forms, but the most common one is DCT-II (Eq. 1). The coefficients are computed according to Eq. 2.

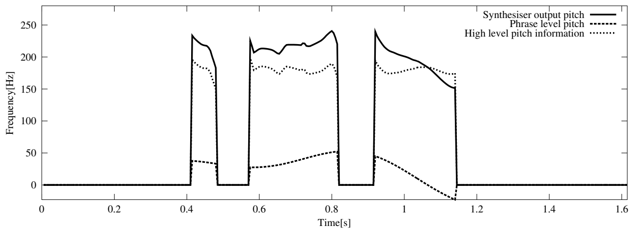

**Fig. 1** An example of a pitch contour for the utterance "Ce mai faci" ("How are you"), the phrase level contour based on the inverse DCT of DCT1-DCT7 coefficients and the high level pitch information.

$$X_k = \frac{1}{2}x_0 + \sum_{n=0}^{N-1} x_n cos\left[\frac{\pi}{N}(n+\frac{1}{2})k\right], k = 0, 1...N-1 \quad (1)$$

$$c_n = \sum_{x=0}^{M-1} s(x) cos\left[\frac{x}{M} n(x+\frac{1}{2})\right] \quad (2)$$

DCT applied to pitch parameterisation has been extensively studied in [7], [13] and [19]. These works prove that DCT is an efficient way to parameterise the pitch with minimum error. Also, the principle behind DCT adheres to the superpositional aspect [14] of the fundamental frequency. The principle states that the pitch can be broken down into separate layers of realisation, heuristically named phrase, word, syllable and phoneme, in the sense that the cognitive process of speech derives a phrase contour unto which the rest of the layers are overlapped. Another important aspect of the DCT is its direct inverse transform. This is needed in the re-synthesis of the pitch contour from the DCT coefficients (Eq. 1).

The method we propose addresses the issue of modelling the phrase level intonation, or trend. Starting from a flat intonation, we would like to derive more dynamic and expressive contours. Therefore, we consider the phrase layer to be represented by the inverse DCT transform of the DCT1 to DCT7 coefficients of the pitch DCT. This assumption is also supported by the results presented in [19]. DCT0 represents the mean of the curve and in our case it is speaker dependent. Using DCT0 in the genome encoding would undesirably change the pitch of the speaker, our focus being on the overall trend of the phrase intonation. The phrase level is then subtracted from the overall contour, and the result is retained and will be referred to as *high level pitch information*. Fig. 1 presents an example of a pitch contour, the phrase level contour based on the inverse DCT of the DCT1-DCT7 coefficients and the high level pitch information. It can be observed that the phrase level contour represents the relative trend of the voiced segments intonation, while the high level information has a relatively flat contour with variations given by the word, syllable and phoneme levels.

Because DCT cannot parameterise fast variations with a small number of coefficients, the unvoiced segments of the F0 contour were interpolated using a cubic function (Eq. 3). During the interactive step we apply the inverse DCT transform over the winner's genome, add the high level pitch information and synthesise the speech using the resulted F0 contour.

$$f(x) = ax^3 + bx^2 + cx + d \tag{3}$$

## 3 Optimisation using CMA-ES

CMA-ES (Covariance Matrix Adaptation - Evolution Strategy) was proposed by Hansen and Ostermeier [5] as an evolutionary algorithm to solve unconstrained or bounded constraint, non-linear optimisation problems defined in a continuous domain. In an evolutionary algorithm, a *population* of genetic representations of the solution space, called *individuals*, is updated over a series of iterations, called *generations*. At each generation, the best individuals are selected as *parents* for the next generation. The function used to evaluate individuals is called the *fitness* function.

The search space is explored according to the genetic operations used to update the individuals in the parent population and generate new offspring. In the case of evolution strategy (ES), the selection and mutation operators are primarily used, in contrast to the genetic algorithm (GA) proposed by Holland [6], which considers a third operator – crossover. Also, in GA the number of mutated genes per individual is determined by the *mutation probability*, while in ES mutation is applied to all genes, slightly and at random.

If mutation is according to a multivariate normal distribution of mean $m$ and covariance matrix $C$, then CMA-ES is a method to estimate $C$ in order to minimise the search cost (number of evaluations). First, for the mean vector $m \in \mathbb{R}^n$, which is assimilated to the preferred solution, new individuals are sampled according to the normal distribution described by $C \in \mathbb{R}^{n \times n}$:

$$\begin{gathered} x_i = m + \sigma y_i \\ y_i \sim N_i(0, C), i = 1..\lambda \end{gathered} \tag{4}$$

where $\lambda$ is the size of the offspring population and $\sigma \in \mathbb{R}_+$ is the step size.

Second, sampled individuals are evaluated using the defined fitness function and the new population is selected. There are two widely used strategies for selection: $(\mu + \lambda)$-ES and $(\mu, \lambda)$-ES, where $\mu$ represents the size of the parent population. In $(\mu + \lambda)$-ES, to keep the population constant, the $\lambda$ worst individuals are discarded after the sampling process. In $(\mu, \lambda)$-ES all the parent individuals are discarded from the new population in favour of the $\lambda$ new offspring.

Third, $m$, $C$ and $\sigma$ are updated. In the case of $(\mu, \lambda)$-ES, which is the strategy we chose to implement our solution, the new mean is calculated as follows:

$$m = \sum_{i=1}^{\mu} w_i x_i \tag{5}$$

$$w_1 \geq .. \geq w_\mu, \sum_{i=1}^{\mu} w_i = 1$$

where $x_i$ is the $i$-th ranked solution vector $(f(x_1) \leq .. \leq f(x_\lambda))$ and $w_i$ is the weight for sample $x_i$.

The covariance matrix $C$ determines the shape of the distribution ellipsoid and it is updated to increase the likelihood of previously successful steps. Details about updating $C$ and $\sigma$ can be found in [4] .

CMA-ES is the proposed solution for Issues #2, #3 and #4 through the generation of several individuals (i.e. speech samples) the user can chose from, the extension of the coefficients' space and the subjective fitness function for the interactive step.

## 4 Proposed solution

Combining the potential of the DCT parameterisation and evolution strategies, we introduce an interactive solution for the intonation optimisation problem, which requires no previous specific knowledge of speech technology. To achieve this, three problems need to be solved: *1)* generate relevant synthetic speech samples for a user to chose from, *2)* minimise user fatigue and *3)* apply the user feedback to improve the intonation of the utterance.

We solve the first problem by using CMA-ES to generate different speech samples, normally distributed around the baseline output of a Romanian speech synthesis system [16] based on HTS (Hidden Markov Models Speech Synthesis System) [21]. We consider a *genome* encoded using a vector of 7 genes, where each gene stores the value of a DCT coefficient, from DCT1 to DCT7. We start with an initial mean vector $m$ that stores the DCT coefficients of the F0 phrase level generated by the HTS system and an initial covariance matrix $C = I \in \mathbb{R}^{7\times 7}$. In each generation, new individuals are sampled according to Eq. (4).

In the next step, the user needs to evaluate generated individuals. If the population size is too large, the user may get tired before a suitable individual is found or might not spot significant differences between the individuals. On the other hand, if the population size is too small and the search space is not properly explored, a suitable individual may not be found. CMA-ES is known to converge faster even with smaller population than other evolutionary algorithms, but it was not previously applied to solve interactive problems. On the other hand, interactive genetic algorithms (IGA) have been extensively studied, but do not converge as fast as CMA-ES for non-linear non-convex problems. Faster convergence means fewer evaluations, therefore reducing user fatigue.

For the interactive version of CMA-ES, we used a *single elimination tournament* fitness [12]. In this case, the individuals are paired at random and play one game per

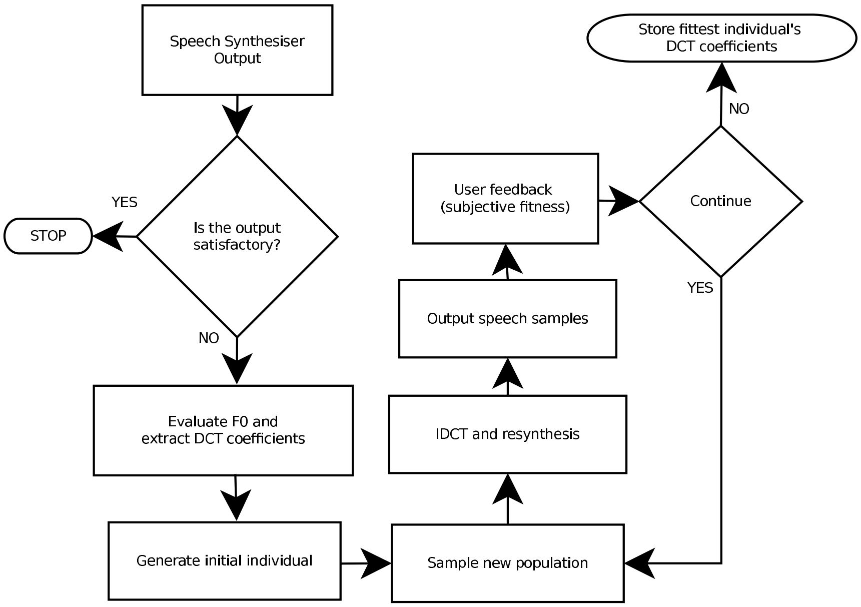
Speech Synthesiser Output
Is the output satisfactory?
YES
STOP
NO
Evaluate F0 and extract DCT coefficients
Generate initial individual
Sample new population
IDCT and resynthesis
Output speech samples
User feedback (subjective fitness)
Continue
NO
YES
Store fittest individual's DCT coefficients

**Fig. 2** Proposed method flow chart

pair. Losers of the game are eliminated from the tournament. The process repeats until a single champion is left. The fitness value of each individual is equal to the number of played games. Each pair of individuals is presented to the user in the form of two speech samples. Being a subjective evaluation, the choice would best suit the user's requirements, thus giving the winner of a population.

The fitness value is used by CMA-ES to update mean vector $m$, the covariance matrix $C$ and the standard deviation $\sigma$. A new population of individuals is sampled based on the updated values and the process repeats. The flow chart of the proposed method is presented in Fig. 2.

## 5 Results

The results presented below focus on establishing the correct scenario for the interactive application and on the ease of use on behalf of the listeners/users. This implies the evaluation of several parameters involved, such as: *initial standard deviation of the population* – gives the amount of dynamic expansion of pitch –, *the population size* – determines the number of samples the user has to evaluate in each generation, *the expressivity and naturalness of the generated individuals* – assures correct values for the pitch contour.

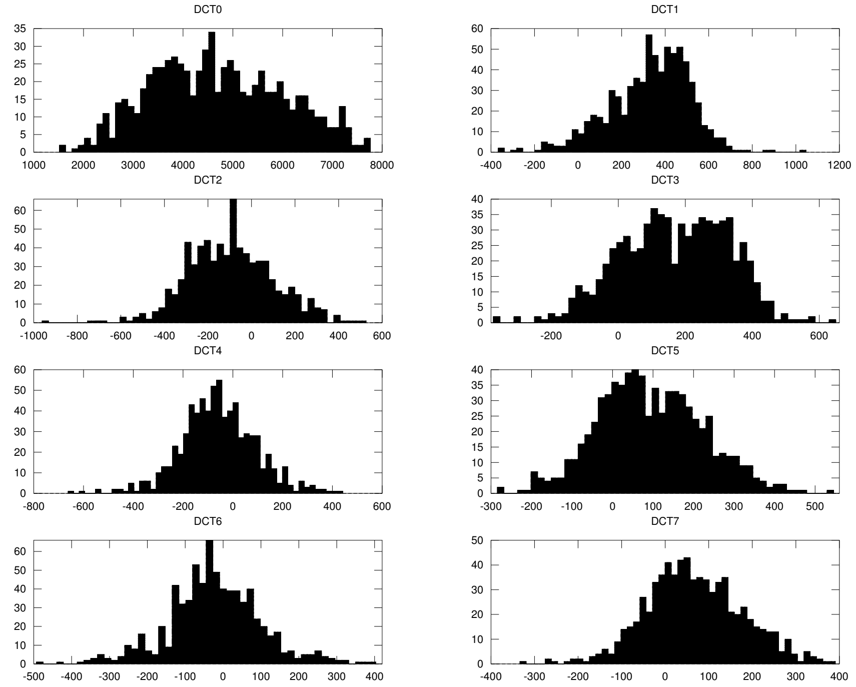

**Fig. 3** The histograms of the first 8 DCT coefficients of the rnd1 subset of the RSS speech corpus. The 0*x* axis represents the values of the DCT coefficients separated in 50 equally spaced bins. The 0*y* axis is the number of coefficients equal to the values within the domain bin.

As a preliminary step in defining the standard deviation of the population, we employed an analysis of all the DCT coefficients within the *rnd1* subset of the Romanian Speech Synthesis corpus [16]. *rnd1* comprises 500 newspaper sentences read by a native Romanian female speaker. The number of phrases within this subset is 730 with an average length of 1.7 seconds. The intonation of the speech is flat, declarative. The histograms of the first 8 DCT coefficients of the phrases in *rnd1* are presented in Fig. 3. We included DCT0 as well for an overall view as it represents the mean of the pitch contour and it is speaker dependent. This coefficient was not used in the estimation of the phrase level contour. The means and standard deviations of the coefficients are presented in Table 1. The average pitch contour resulted from the mean values of the DCT coefficients and the average duration of the *rnd1* subset is shown in Fig. 4.

DCT1 has the most important influence in the F0 contour after DCT0. The mean value of the DCT1 coefficient is 331.75 with a standard deviation of 185.85 and the maximum F0 variation is given by the *+1 std. dev.* (i.e. 331.75+185.85 = 517.6) of around 40 Hz. One of the issues addressed in this paper is the expansion of the pitch range. This means that having a standard deviation of the flat intonation

**Table 1** Means and standard deviation of the DCT coefficients in *rnd1* subset with corresponding variations in Hz for an average length of 1.7 seconds.

| ’Coefficient | Mean | Mean F0 [Hz] | Standard deviation | Maximum F0 deviation [Hz] | |
|---|---|---|---|---|---|
| | | | | - 1 std dev | +1 std dev |
| DCT0 | 4690.300 | 251-257 | 1318.300 | 179-186 | 322-329 |
| DCT1 | 331.750 | ± 4 | 185.850 | ±12 | ±40 |
| DCT2 | -95.087 | ±7 | 197.470 | ±22 | ±7 |
| DCT3 | 168.270 | ±12 | 161.030 | ±0.55 | ±25 |
| DCT4 | -57.100 | ±4 | 151.600 | ±16 | ±7 |
| DCT5 | 94.427 | ±7 | 130.150 | ±2 | ±17 |
| DCT6 | -22.312 | ±1 | 123.020 | ±11 | ±7 |
| DCT7 | 67.095 | ±5 | 110.370 | ±3 | ±13 |

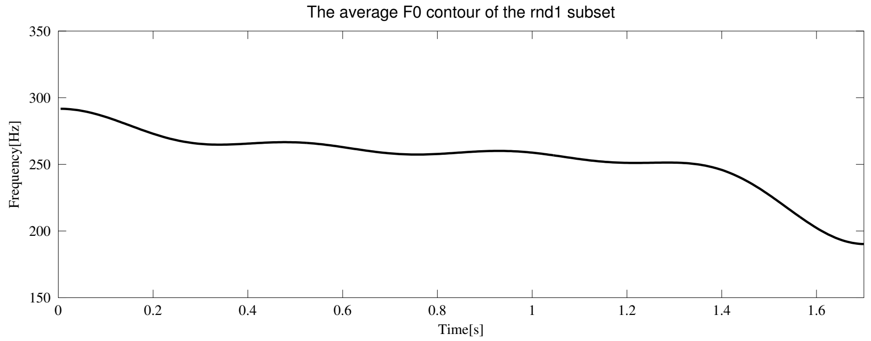

**Fig. 4** The average pitch contour resulted from the mean values of the DCT0-DCT7 coefficients for the average length of 1.7 seconds in the *rnd1* subset.

speech corpus, we should impose a higher value for it while generating new speech samples, but it should not go up to the point where the generated pitch contours contain F0 values which are not natural. In Fig. 5 we compare the third generation for an initial standard deviation of 150 and 350 respectively. We can observe in the 350 case that individual 3 has F0 values going as low as 50 Hz – unnatural, while for a standard deviation of 150, the F0 contours do not vary too much from the original one and lead to a less dynamic output. Given these results, we selected a standard deviation of 250. An important aspect to be noticed from Table 1 is that all the 7 coefficients have approximately the same standard deviation. This means that imposing a variation based on DCT1 does not exceed natural values for the rest of the coefficients.

The single elimination tournament fitness we used to evaluate the individuals requires the user to provide feedback for $n-1$ games, where $n$ is the population size. So that the population size has a great importance in setting up the interactive

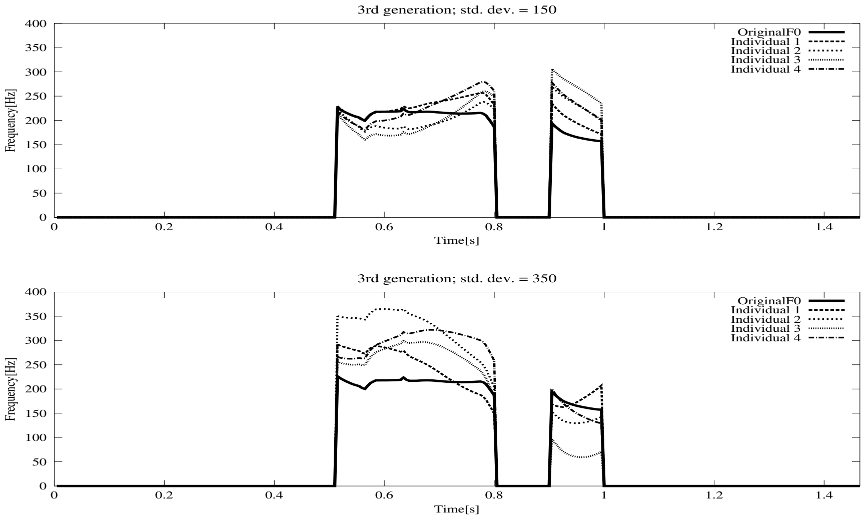

**Fig. 5** The 3rd generation population of the F0 contour for the phrase "Ce mai faci?" ("How are you?"), with an initial standard deviation of 150 and 350 respectively. Original F0 represents the pitch contour produced by the synthesiser.

application. Several values have been selected for it and the results are shown in Fig. 6. Although the highest the number of individuals the more samples the user can choose from, this is not necessarily a good thing in the context of user fatigue. But having only 2 individuals does not offer enough options for the user to choose from. We therefore suggest the use of 4 individuals per generation as a compromise between sample variability and user fatigue.

Another evaluation is the observation of the modification of the pitch contour from one generation to the other. Fig. 7 presents the variation of F0 from the initial population to the third. It can be observed that starting with a rather flat contour, by the third generation the dynamics of the pitch are much more expanded, resulting a higher intonation variability within and between generations. It is also interesting to observe the phrase level contours (Fig. 8). This is a more relevant evaluation as it shows the different trends generated by CMA-ES and the trend selected by the user in each generation. The selected trend can be used in the adaptation of the overall synthesis. In our example, the user selected an intonation with a high starting point and a descending slope afterwards, while another user could have chosen individual 1 which contains an initial ascending slope.

In order to establish the naturalness of the generated individuals and the enhanced expressivity of the winners of each generation, a small listening test was conducted. At first, a user was asked to select the winners over 4 generations for 10 phrases. Initial standard deviation was 250 and with a population size of 4. Then 10 listeners had to attribute Mean Opinion Scores (MOS) for the samples in two categories: *Naturalness* – the generated samples were compared to original recordings on a

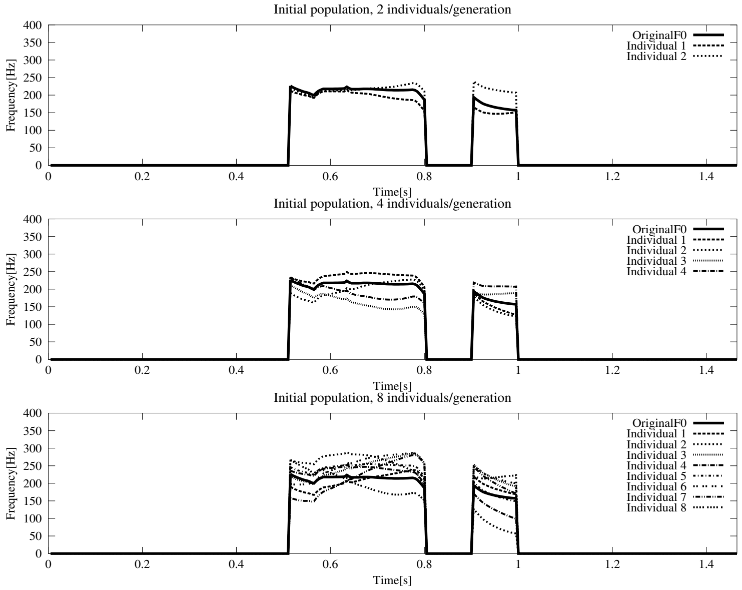

**Fig. 6** Variation in the population size. Phrase ”Ce mai faci?” (”How are you?”). Original F0 represents the pitch contour produced by the synthesiser.

scale of [1 - Unnatural] to [5 - Natural]. All the individuals of the four generations were presented. *Expressivity* – the winners of each generation were compared to the correspondent synthesised versions of them. The listeners had to mark on a scale of [1-Less expressive] to [5-More expressive] the generated samples in comparison to the synthesiser's output. The results of the test are presented in Fig. 9. In the naturalness test, all the generations achieved a relatively high MOS score, with some minor differences for the $4^{th}$ generation. The expressivity test reveals the fact that all the winning samples are more expressive than the originally synthesised one. The test preliminary conclude the advantages of this method. While maintaining the naturalness of the speech, its expressivity is enhanced.

Examples of speech samples generated by our method can be found at `http://www.romaniantts.com/nicso2011`.

## 6 Conclusions

We introduced a new method for intonation optimisation of a speech synthesis system based on CMA-ES and DCT parameterisation of the pitch contour. The interactive manner of the optimisation allows the users to select an output which best suits their expectations. The novelty of the solution consists in using no prosodic annotations of the text, no deterministic rules and no predefined speaking styles. Also,

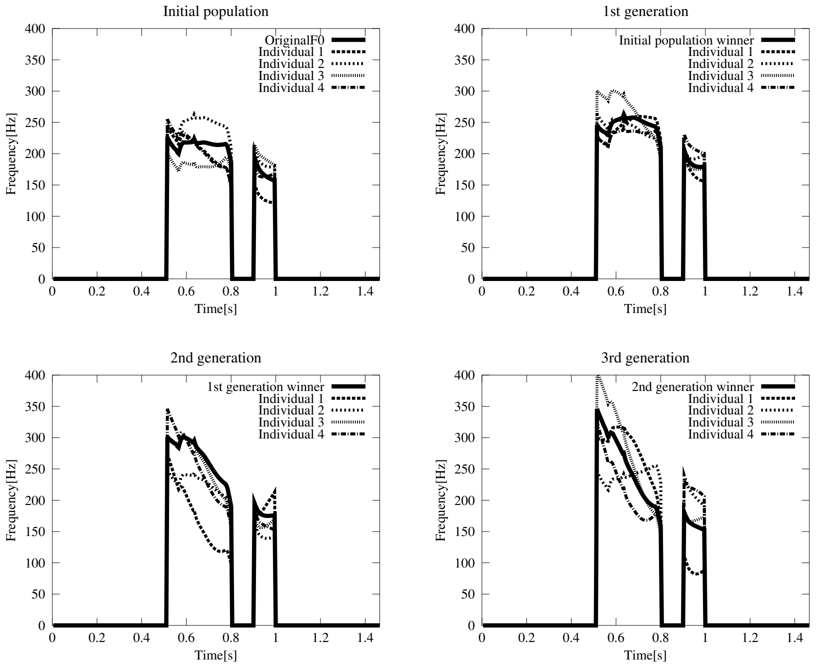

**Fig. 7** Evolution of the F0 contour over 3 generations, standard deviation = 250, phrase ”Ce mai faci?” (”How are you?”). Original F0 represents the pitch contour produced by the synthesiser.

to the best of our knowledge, this is one of the first applications of CMA-ES for an interactive problem.

The evaluation of the system's parameters provide the guidelines of the setup for an interactive application. The proposed solutions ensure an optimal value for standard deviation and population size in order to concurrently maintain the naturalness of the speech samples, while expanding the dynamics of the pitch. The latter indicators have been evaluated in the listening test. The listening test also determined the enhancement of the expressivity of the samples.

One drawback to our solution is the lack of individual manipulation of each of the 7 DCT coefficients in the genome, unattainable in the context of the evolutionary algorithm chosen. However the coefficients' statistics showed that the average standard deviation is similar and thus the choice for the initial standard deviation does not alter the higher order coefficients.

As the results obtained in this preliminary research have achieved a high-level of intonational variation and user satisfaction, a web-based application of the interactive optimisation is under-way. The application would allow the user to select the entire utterance or just parts of it – i.e., phrases, words or even syllables – for the optimisation process to enhance. For a full prosodic optimisation, we would like to include the duration of the utterance in the interactive application as well.

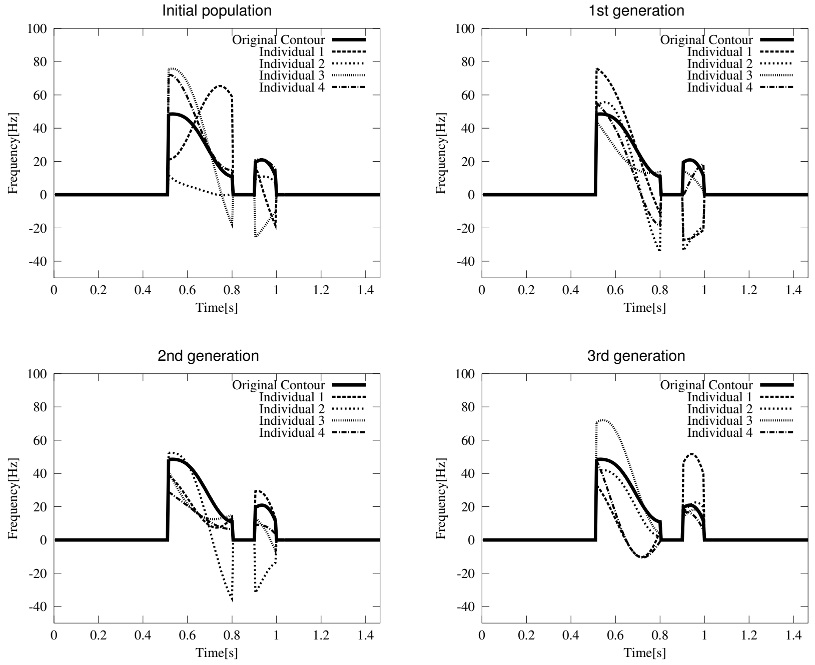

**Fig. 8** Evolution of the phrase contour trend over 3 generations for the utterance ”Ce mai faci” (”How are you”). Original contour represents the pitch contour produced by the synthesiser.

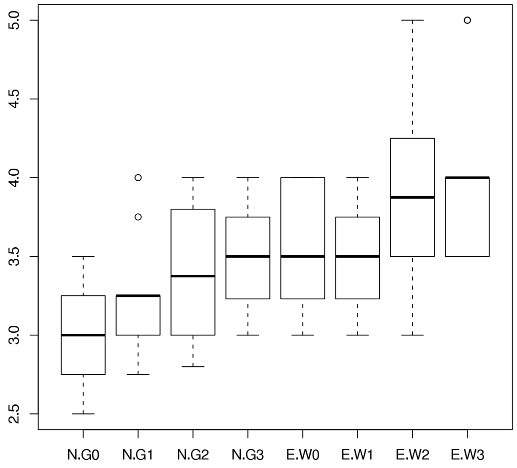

**Fig. 9** Box plots results of the listening test. N-G*x* represent the results for the naturalness test of each generation and E-G*x* represent the results for the expressivity test of each generation. The median is represented by a solid bar across a box showing the quartiles and whiskers extend to 1.5 times the inter-quartile range.

One interesting development would be a user-adaptive speech synthesiser. Based on previous optimisation choices, the system could adapt in time to a certain prosodic realisation. Having set up the entire workflow, testing different types of fitness functions is also of great interest.

## 7 Acknowledgements

This work has been funded by the European Social Fund, project POSDRU 6/1.5/S/5 and the national project TE 252 financed by CNCSIS-UEFISCSU.

## References

[1] D'Este, F., Bakker, E.: Articulatory Speech Synthesis with Parallel Multi-Objective Genetic Algorithms. In: Proc. ASCI (2010)

[2] Fujisaki, H., Ohno, S.: The use of a generative model of F0 contours for multilingual speech synthesis. In: ICSLP-1998, pp. 714–717 (1998)

[3] Fukumoto, M.: Interactive Evolutionary Computation Utilizing Subjective Evaluation and Physiological Information as Evaluation Value. In: Systems Man and Cybernetics, pp. 2874 – 2879 (2010)

[4] Hansen, N.: The CMA evolution strategy: A tutorial. Tech. rep., TU Berlin, ETH Zurich (2005)

[5] Hansen, N., Ostermeier, A.: Adapting arbitrary normal mutation distributions in evolution strategies: the covariance matrix adaptation. In: Proceedings of IEEE International Conference on Evolutionary Computation, pp. 312 –317 (1996)

[6] Holland, H.: Adaptation in Natural and Artificial Systems. University of Michigan Press (1975)

[7] Latorre, J., Akamine, M.: Multilevel Parametric-Base F0 Model for Speech Synthesis. In: Proc. Interspeech (2008)

[8] Lv, S., Wang, S., Wang, X.: Emotional speech synthesis by XML file using interactive genetic algorithms. In: GEC Summit, pp. 907–910 (2009)

[9] Marques, V.M., Reis, C., Machado, J.A.T.: Interactive Evolutionary Computation in Music. In: Systems Man and Cybernetics, pp. 3501–3507 (2010)

[10] McDermott, J., O'Neill, M., Griffith, N.J.L.: Interactive EC control of synthesized timbre. Evolutionary Computation **18**, 277–303 (2010)

[11] Moisa, T., Ontanu, D., Dediu, A.: Speech synthesis using neural networks trained by an evolutionary algorithm. In: Computational Science - ICCS 2001, *Lecture Notes in Computer Science*, vol. 2074, pp. 419–428. Springer Berlin / Heidelberg (2001)

[12] Panait, L., Luke, S.: A comparison of two competitive fitness functions. In: Proceedings of the Genetic and Evolutionary Computation Conference, GECCO '02, pp. 503–511 (2002)

[13] Qian, Y., Wu, Z., Soong, F.: Improved Prosody Generation by Maximizing Joint Likelihood of State and Longer Units. In: Proc. ICASSP (2009)

[14] Sakai, S.: Additive modelling of English F0 contour for Speech Synthesis. In: Proc. ICASSP (2005)

[15] Silverman, K., Beckman, M., Pitrelli, J., Ostendorf, M., Wightman, C., Price, P., Pierrehumbert, J., Hirschberg, J.: ToBI: A standard for labeling English prosody. In: ICSLP-1992, vol. 2, pp. 867–870 (1992)

[16] Stan, A., Yamagishi, J., King, S., Aylett, M.: The Romanian speech synthesis (RSS) corpus: Building a high quality HMM-based speech synthesis system using a high sampling rate. Speech Communication **53**(3), 442 – 450 (2011). DOI DOI:10.1016/j.specom.2010.12.002

[17] Tao, J., Kang, Y., Li, A.: Prosody conversion from neutral speech to emotional speech. IEEE Trans. on Audio Speech and Language Processing **14**(4), 1145–1154 (2006). DOI {10.1109/TASL.2006.876113}

[18] Taylor, P.: The tilt intonation model. In: ICSLP-1998, pp. 1383–1386 (1998)

[19] Teutenberg, J., Wilson, C., Riddle, P.: Modelling and Synthesising F0 Contours with the Discrete Cosine Transform. In: Proc. ICASSP (2008)

[20] Yamagishi, J., Onishi, K., Masuko, T., Kobayashi, T.: Acoustic modeling of speaking styles and emotional expressions in hmm-based speech synthesis. IEICE - Trans. Inf. Syst. **E88-D**, 502–509 (2005)

[21] Zen, H., Nose, T., Yamagishi, J., Sako, S., Tokuda, K.: The HMM-based speech synthesis system (HTS) version 2.0. In: Proc. of Sixth ISCA Workshop on Speech Synthesis, pp. 294–299 (2007)