Achieving Shorter Search Times in Voice Conversion Using Interactive Evolution

Yuji Sato

Faculty of Computer and Information Sciences, Hosei University
3-7-2 Kajino-cho, Koganei-shi, Tokyo 184-8584, Japan
yuji@k.hosei.ac.jp

Abstract. We have already proposed using evolutionary computation to adjust the voice quality conversion parameters, and we have reported that this approach produces results that are not only closer to the desired target than the results of parameter adjustment based on designer experience or trial and error, but which also have relatively little sound quality degradation. In this paper we propose improved techniques for the generation of initial entities and genetic manipulation in order to reducing the workload associated with human evaluation in interactive evolution. We perform voice quality conversion experiments both on natural speech recorded with a microphone and on synthetic speech generated from text data. As a result, we confirm that the proposed improvements make it possible to perform voice quality conversion more efficiently than when using the technique proposed earlier.

1 Background and Basic Idea for Reducing the Workload

New markets are appearing using voice quality conversion technology. These include multimedia-content editing, computer games, and man-personal machine interfaces. For example, in multimedia-content editing, adding narration to business content such as presentation material and digital catalogs or to personal content such as photo albums and self-produced video can enhance content. It is not necessarily easy, however, for the general user to provide narration in a clear and intelligible voice, and the need for voice quality conversion can be felt here. Against the above background, we have proposed the application of evolutionary computation to parameter adjustment for the sake of voice quality conversion, and it has also been shown that the use of evolutionary computation for parameter adjustments can be effective at improving the clarity not only of natural speech that has been subjected to voice conversion but also of synthetic speech generated automatically from text data.

In the system we proposed earlier [1], we represented the three variables α, β and γ as real numbers, and we defined a chromosome as an array of the form $[\alpha, \beta, \gamma]$. Then, we performed the crossover operation by randomly selecting one variable from among the three array elements and swapping the values of this variable between two parent entities. Next, we use mutation as represented by Eq. (1) to raise the probability that target mutants are in the vicinity of parents

K. Deb et al. (Eds.): GECCO 2004, LNCS 3103, pp. 1328–1329, 2004.

and to improve local searching. In the equation, C_i represents a modification coefficient for generation i, I is a unit matrix, k is a constant, and N is a normal distribution function with a mean vector of 0 and a covariance of kI and is common to all elements.

$$C_{i+1} = C_i + N(0, kI) \tag{1}$$

Here, the first problem to be addressed in practical implementations of the above system is that the interactive evolution scheme places a heavy workload on the evaluators. At first, the evaluator workload depends on the number of evaluations that have to be performed. We think that if empirical conversion coefficients for different conversion objectives are used as the starting values, then it should be possible to reduce the total number of entities even if there are no empirical conversion coefficients available from previous manual searches.

Next, it is thought that the search performance can be improved by making changes to the crossover and spontaneous mutation operations so as to reduce the number of generations needed to arrive at a practical quasi-optimal solution. In new operation, one of the two entities generated by the crossover operation shown the above is randomly subjected to crossovers in which the average value of each coefficient in the two parent entities [2] are obtained.

For spontaneous mutations, the standard deviation of the mutation distribution was set small as shown in Eq. (2) for entities where crossovers were performed just by swapping coefficients as in the conventional approach.

$$C_{i+1} = C_i + N(0, 0.000025I) \tag{2}$$

Conversely, a larger standard deviation was set for entities where crossovers were performed by taking the average of two coefficients as shown in Eq. (3).

$$C_{i+1} = C_i + N(0, 0.01I) \tag{3}$$

That is, the emphasis is placed on local search performance for entities where crossovers are performed in the same way as in earlier systems, and the emphasis is placed on increasing diversity and searching new spaces for entities where crossovers are performed by obtaining the average of two coefficients.

Using an evolutionary computation conversion method reported earlier [1] and the conversion technique proposed here, we compared the number of evaluations (the number of generations multiplied by the number of entities). In terms of the number of evaluations, convergence is achieved with an even smaller number of evaluations (from 100 times to 40 times) when the measures proposed here are applied.

References

1. Sato, Y.: Voice Conversion Using Interactive Evolution of Prosodic Control. In: Proc. of the 2002 Genetic and Evolutionary Computation Conference (GECCO-2002), Morgan Kaufmann Publishers, San Francisco, CA (2002) 1204–1211
2. Bäck, T., Fogel, D.B., and Michalewicz, Z. (eds.): Evolutionary Computation 1: Basic Algorithms and Operators. Institute of Physics Publishing, Bristol, UK (2000)