Voice Conversion Using Interactive Evolution of Prosodic Control

Yuji Sato

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

Abstract

This paper proposes the application of evolutionary computation, a stochastic search technique that parallels the evolution of living organisms, to parameter adjustment for voice conversion, and reports on several experimental results applicable to the fitting of prosodic coefficients. Here, because of the difficulty involved in providing a clear fitness function for evaluating evolutionary computation, we adopt a system of interactive evolution in which genetic manipulation is repeated while evaluation is performed subjectively based on human feelings. It was found that the use of evolutionary computation achieves voice conversion closer to the target in question than parameter adjustment based on designer experience or trial and error, and that degradation in sound quality is relatively small giving no impression of a processed voice.

1 INTRODUCTION

With the coming of the multimedia era, the market for multimedia information devices centered about personal computers is experiencing rapid growth. Likewise, the market for multimedia application software is taking off giving rise to an environment in which users can manipulate images and sound with ease. In particular, speech synthesis technology is expected to generate a large market for a wide rage of applications from the reading of E-mail and text data on the World Wide Web to the speaking of road traffic reports provided by navigation devices. Nevertheless, mechanically synthesized speech by a rule-based speech synthesis system or similar suffers from a variety of problems. These include an impression of discontinuity between phoneme fragments, degraded sound quality due to repeated signal processing, and limitations in sound-source/articulation segregation models. In other words, the synthesis of natural speech is extremely difficult. Current technology tends to produce mechanical or unintelligible speech, and problems such as these are simply delaying the spread of speech synthesis products.

Research has also begun on the application of voice processing to narration when editing multimedia content as in a spoken presentation. The need for voice conversion (processing) arises from the fact that most people have difficulty speaking with an expressive and clear voice. However, only qualitative know-how has so far been obtained in the development of voice-processing technology for converting original speech to clear narration. Parameter setting is currently performed on a trial and error basis making adjustments difficult.

Against the above background, this research aims to establish technology for converting original human speech or speech mechanically synthesized from text to clear speech rich in prosodic stress. As the first step to this end, we have proposed the application of evolutionary computation to parameter adjustment for the sake of voice conversion using original speech recorded by a microphone as input data, and have reported on several experimental results applicable to the fitting of prosodic coefficients [Sato 1997]. In this paper, we show that parameter adjustment using evolutionary computation can be effective not only for voice conversion using original speech as input but also for improving the clarity of speech mechanically synthesized from text. We also investigate why parameter adjustment using evolutionary computation is more effective than that based on trial and error by an experienced designer.

2 VOICE ELEMENTS AND VOICE CONVERSION

This section summarizes the feature quantities needed for voice conversion and describes voice conversion by prosodic control.

2.1 VOICE ELEMENTS

In human speech production, the vocal cords serve as the sound generator. The vocal cords, which are a highly flexible type of muscle located deep in the throat, are made to vibrate by breath expelled from the lungs, thereby causing acoustic vibrations in the air (sound waves). The waveform of this acoustic signal is approximately triangular or saw-tooth in form and consists of harmonic components that are integer multiples of the fundamental frequency of the sound wave. This acoustic signal that has a broad range of harmonic components of a constant interval propagates through the vocal tract from the vocal cords to the lips and acquires resonances that depend on the shape of the vocal tract. This transformations results in the production of phonemes such as /a/ or /i/, which are finally emitted from the lips as speech. That is to say, the human voice characteristics are determined by three factors: sound generation, propagation in the vocal tract, and emission. The vocal cords control the pitch of the voice and the shape of the vocal tract controls prosody. If we define voice quality in terms of properties such as timbre, we can consider voice quality to be determined by both the state of the vocal cords and the state of the vocal tract [Klatt 1990]. That is to say, we can consider pitch structure, amplitude structure, temporal structure and spectral structure as the feature quantities for the control of voice quality.

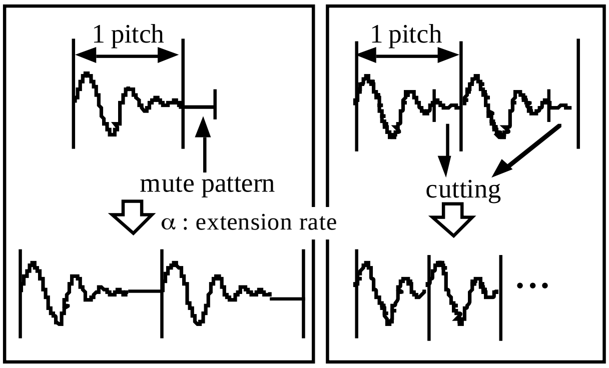
2.2 MODIFICATION OF VOICE QUALITY THROUGH PROSODIC ADJUSTMENT

Research on the features of the voices of professional announcers has clarified to some extent the qualitative tendencies that are related to highly-intelligible speech. It is known, for example, that raising the overall pitch slightly and increasing the acoustic power of consonants slightly increases intelligibility [Kitahara 1992]. It remains unclear, however, to what specific values those parameters should be set. Moreover, it is generally difficult to control dynamic spectral characteristics in real time. In other words, it is difficult to even consider adjusting all of the control parameters to begin with. Therefore, sought to achieve voice conversion by limiting the data to be controlled to pitch data, amplitude data, and temporal structure prosodic data**.**

The pitch conversion method is shown in Fig. 1. Pitch is raised by cutting out a part of the waveform within one pitch unit. Pitch is lowered by inserting silence into a pitch unit. Modification of the temporal structure is accomplished as illustrated in Fig. 2. The continuation length is accomplished by using the TDHS [Malah 1979] enhancement method to extend or contract the sound length without changing the pitch. Amplitude is modified on a logarithmic power scale according to the formula

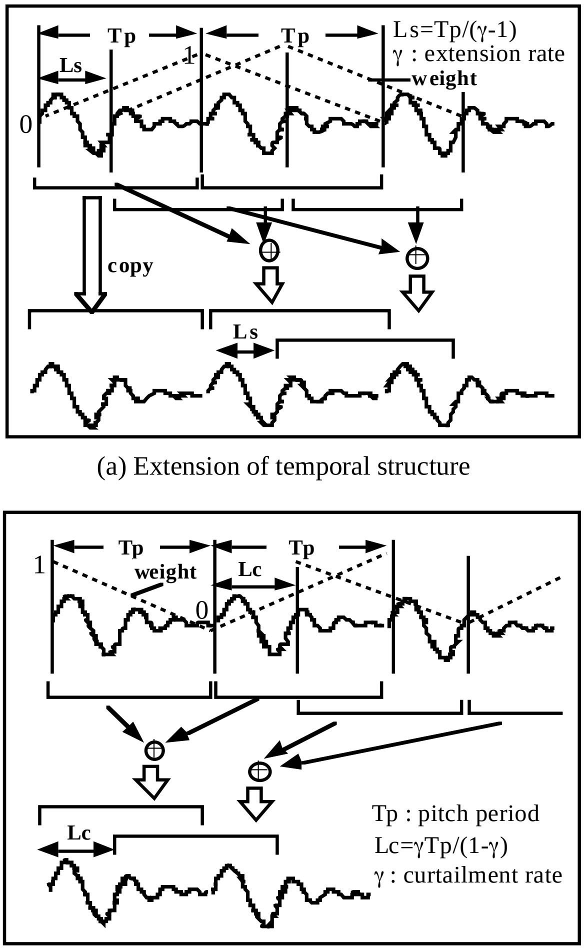
$$\log_{10} W_{i+1}{}^{2} = \log_{10} W_i^2 + \beta \qquad (1)$$

Where W_i is the current value and β is the modification coefficient.



(a) Extension (b) Curtailment

Figure 1: Extension and curtailment of pitch period. Pitch is raised by cutting out a part of the waveform within one pitch unit. Pitch is lowered by inserting silence into a pitch unit..



(a) Extension of temporal structure

(b) Curtailment of temporal structure

Figure 2: Extension and curtailment of temporal structure. The continuation length is accomplished by using the TDHS enhancement method to extend or contract the sound length without changing the pitch.

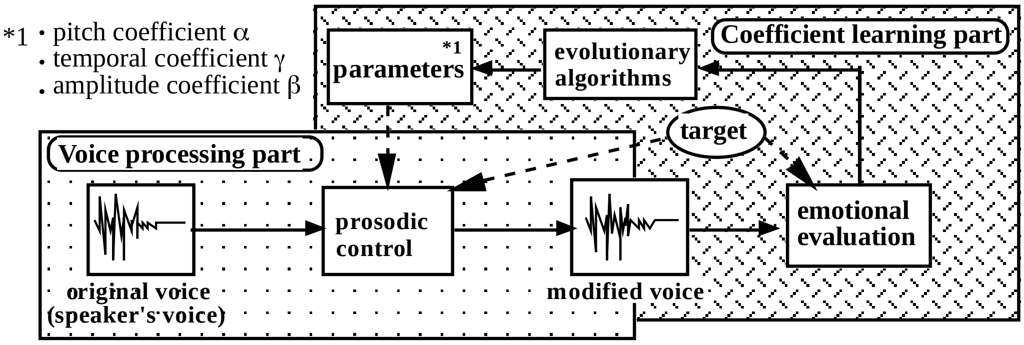
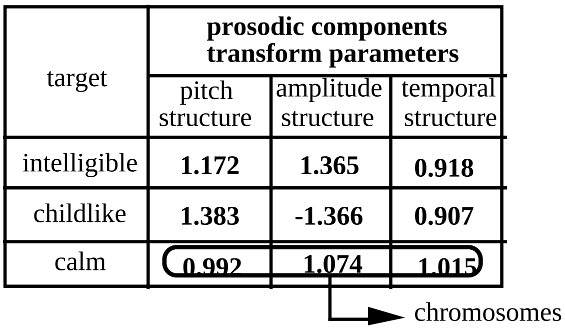


Figure 3: Block diagram of proposed voice quality conversion system. The system comprises a voice processing part and prosody control coefficient learning part.



target	prosodic components transform parameters		
	pitch structure	amplitude structure	temporal structure
intelligible	**1.172**	**1.365**	**0.918**
childlike	**1.383**	**-1.366**	**0.907**
calm	**0.992**	**1.074**	**1.015**

chromosomes

Figure 4: Example of the chromosomes. It is defined by an array, [pitch modification factor α, amplitude modification factor β, continuation time factor γ].

3 PROSODIC COEFFICIENT FITTING BY EVOLUTIONARY COMPUTATION

3.1 CONFIGURATION OF THE VOICE MODIFICATION SYSTEM

The configuration of the voice modification system is illustrated in Fig. 3. The system comprises a voice processing part and prosody control coefficient learning part. The voice modification unit changes voice quality, targeting terms that express emotional feelings, such as “clear,” and “cute.” The modification of prosodic information is done by the prosodic control unit. To prevent degradation of voice quality, the processing is done at the waveform level as described above rather than at the parameter level, as is done in the usual analysis-synthesis systems. The modification coefficient learning unit is provided with qualitative objectives, such as terms of emotion, and the modification coefficients used for prosodic modification targeting those objectives are acquired automatically by learning. As the learning algorithm, this unit employs evolutionary computation, which is generally known as an effective method for solving problems that involve optimization of a large number of combinations.

3.2 OVERVIEW OF INTERACTIVE EVOLUTION OF PROSODIC CONTROL

The first step in this procedure is to define chromosomes, i.e., to substitute the search problem for one of determining an optimum chromosome. As shown in Fig. 4, we define a chromosome as a one-dimensional real-number array corresponding to a voice-conversion target (an emotive term) and consisting of three prosody modification coefficients. Specifically, denoting the pitch modification factor as α, the amplitude modification factor as β, and the continuation time factor as γ, we define a chromosome as the array [α, β, γ]. The next step is to generate individuals. Here, we generate 20, and for half of these, that is, 10 individuals, chromosomes are defined so that their prosody modification coefficients change randomly for each voice-conversion target. For the remaining 10, chromosomes are defined so that their coefficients change randomly only within the vicinity of prosody-modification-coefficient values determined from experience on a trial and error basis. In the following step, evaluation, selection, and genetic manipulation are repeated until satisfactory voice quality for conversion is attained. Several methods of evaluation can be considered here, such as granting points based on human subjectivity or preparing a target speech waveform beforehand and evaluating the mean square difference between this target waveform and the output speech waveform from voice-conversion equipment. In the case of evolutionary computation, a designer will generally define a clear evaluation function beforehand for use in automatic recursion of change from one generation to another. It is difficult to imagine, however, a working format in which an end user himself sets up a clear evaluation function, and in recognition of this difficulty, we adopt a system of interactive evolution [Sims 1991, Takagi 2001] in which people evaluate results subjectively (based on feelings) for each generation.

3.3 GENETIC MANIPULATION

3.3.1 Selection Rule

Culling and selection are based on a fitness value, as shown in Fig. 5. First, the individuals are sorted by their fitness values. In the example shown in Fig. 5, 20 individuals are sorted in order of high fitness value with respect to the objective of high intelligibility. The population is then culled. Here, The half of the individuals with the lowest fitness values is culled. The proportion of the population culled does not have to be 50%; another approach is to cull all individuals whose fitness values are below a certain standard value. Next, the population is replenished by replacing the culled individuals with a new generation of individuals picked by roulette selection [Goldberg 1989] in this example. To produce the new generation, first two chromosomes are selected as the parents. Offspring are generated from the parents by the crossover and mutation process described below. Here, the probability of selecting the two parent chromosomes is proportional to the fitness values. Furthermore, duplication in the selection is permitted. All individuals are parent candidates, including the culled individuals. In other words, taking *M* as the number of individuals to be culled, we randomly select only *M* pairs of individuals from the current generation of *N* individuals (I_1 to I_N), permitting duplication in the selection. The crossover and mutation genetic manipulation operations are performed on those pairs to provide *M* pairs of individuals for replenishing the population. Here, the probability $P(I_i)$ of an individual I_i being selected as a parent for creating the next generation of individuals is determined by the following equation. The term $f(I_i)$ in this equation expresses the degree of adaptability of I_i.

$$P(I_i) = f(I_i) / \left\{ \sum_{j=1}^{N} f(I_j) / N \right\} \qquad (2)$$

Although the method used here is to assign a fitness value to each individual and cull the individuals that have low values, it is also possible to select the individuals to be culled by a tournament system. In that case, we do not have access to the fitness values, so we considered random selection of the parent individuals.

3.3.2 Crossover and Mutation

Figure 6 presents an example of crossover. In the crossover operation, any one column is chosen and the values in that column are swapped in the two parent individuals. In Fig. 6, the modification coefficients for continuation length are exchanged between the two parents. The crossover genetic manipulation has the effect of propagating bit strings (chromosome structural components) that are linked to high fitness values to another individual. If these structural components, which are referred to as building blocks [Goldberg 1989], are successfully assembled in an accurate manner, then an effective search is accomplished.

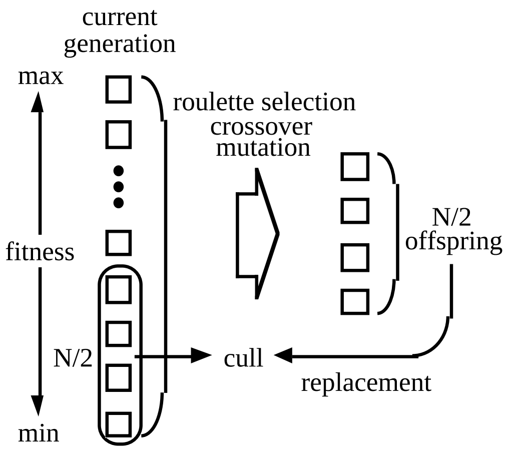
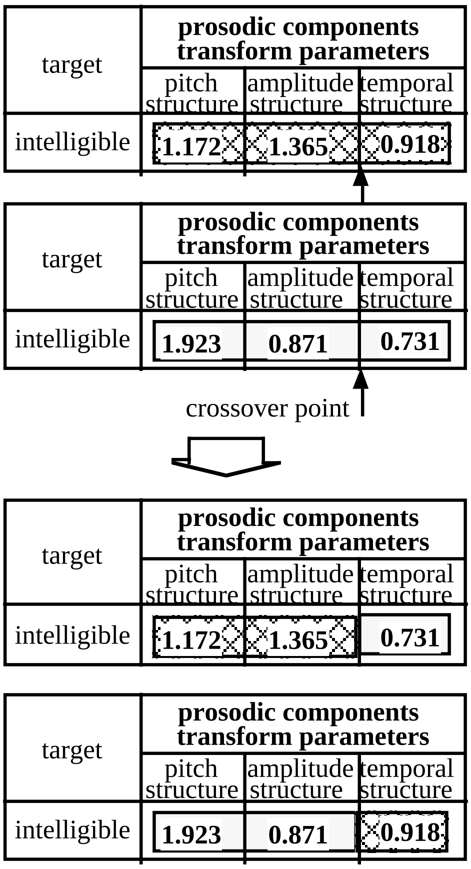


Figure 5: Selection rule.



target	prosodic components transform parameters		
	pitch structure	amplitude structure	temporal structure
intelligible	1.172	1.365	0.918

target	prosodic components transform parameters		
	pitch structure	amplitude structure	temporal structure
intelligible	1.923	0.871	0.731

target	prosodic components transform parameters		
	pitch structure	amplitude structure	temporal structure
intelligible	1.172	1.365	0.731

target	prosodic components transform parameters		
	pitch structure	amplitude structure	temporal structure
intelligible	1.923	0.871	0.918

Figure 6: Example of crossover. In the crossover operation, any one column is chosen and the values in that column are swapped in the two parent individuals.

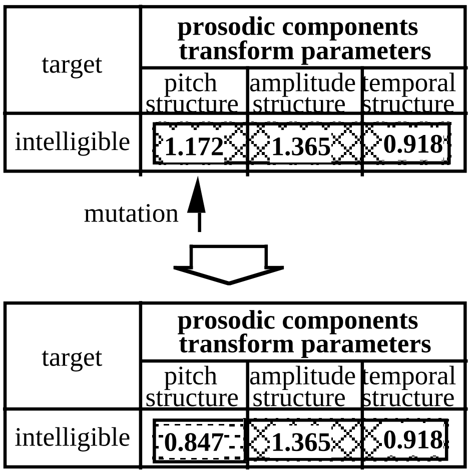


Figure 7: Example of mutation. In this example, the modification parameter for pitch is chosen and the value is varied in the range from 1.172 to 0.847.

Figure 7 shows an example of mutation whereby a prosody modification coefficient is arbitrarily selected and randomly changed. In this example, the operation selects the modification coefficient related to pitch and its value mutates from 1.172 to 0.847. Here, we use mutation as represented by Eq. (3) to raise the probability that target mutants are in the vicinity of parents 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.

$$C_{i+1} = C_i + \delta_i \quad (i = 1 \; to \; 20) \tag{3}$$

$$\delta_i = N(0, kI) \tag{4}$$

This mutation operation has the effects of escaping from local solutions and creating diversity. In addition, crossover and mutation combined raise the fitness value, that is, the vicinity of the modification coefficient can be efficiently searched near the voice-conversion target. Moreover, as multiple individuals are performing parallel searches from different initial values, initial-value dependency is low and positive effects from parallel processing can be expected.

In the experiments described below, we used a crossover rate of 0.5 and a mutation rate of 0.3.

4 EVALUATION EXPERIMENTS

4.1 EXPERIMENT WITH ORIGINAL SPEECH AS INPUT DATA

4.1.1 Voice Stimuli

The original voice sample, *S*0, was the sentence, “Let me tell you about this company.” spoken by a female in Japanese. Five modified samples, *SA*1 through *SA*5, that correspond to the five emotive terms, “intelligible,” “childish,” “joyful,” “calm,” and “angry,” were produced by applying prosody modification coefficients obtained by the evolutionary computation learning scheme described above. In addition, five modified samples, *SB*1 through *SB*5, that correspond to the same five emotive terms, “intelligible,” “childish,” “joyful,” “calm,” and “angry,” were produced by applying prosody modification coefficients obtained by trial and error based on the experience of a designer.

4.1.2 Experimental Method

The subjects of the experiments were 10 randomly selected males and females between the ages of 20 and 30 who were unaware of the purpose of the experiment. Voice sample pairs *S*0 together with *SAi* (*i* = 1 to 5) and *S*0 together with *SBi* (*i* = 1 to 5) were presented to the test subjects through speakers. The subjects were instructed to judge for each sample pair whether voice modification corresponding to the five emotive terms specified above had been done by selecting one of three responses: “Close to the target expressed by the emotive term,” “Can't say,” and “Very unlike the target.” To allow quantitative comparison, we evaluated the degree of attainment (how close the modification came to the target) and the degree of good or bad impression of the sample pairs on a nine-point scale for the childish emotive classification. Subjects were allowed to hear each sample pair multiple times.

4.1.3 Experimental Results

The results of the judgments of all subjects for voice sample pairs *S*0 - *SAi* (*i* = 1 to 5) and *S*0 - *SBi* (*i* = 1 to 5) are presented in Fig. 8 as a histogram for the responses “Close to the target” and “Very unlike the target”. From those results, we can see that although the trial and error approach to obtaining the modification coefficients was successful for the “childish”, “intelligible”, and “joyful” classifications, the modification results were judged to be rather unlike the target for the “calm” and “angry” classifications. In contrast to those results, the samples produced using the modification coefficients obtained by the evolutionary computation approach were all judged to be close to the target on the average.

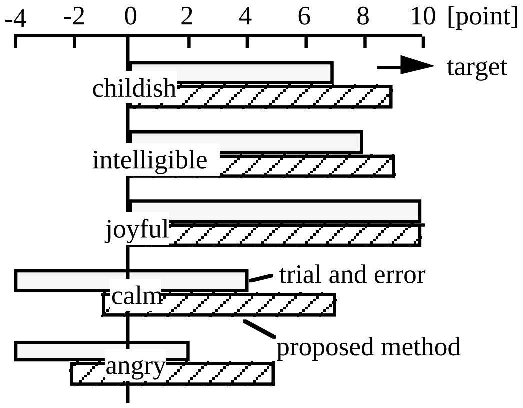


Figure 8: The results of the judgments of all subjects for voice sample pairs. The results are presented as a histogram for the responses "Close to the target" and "Very unlike the target".

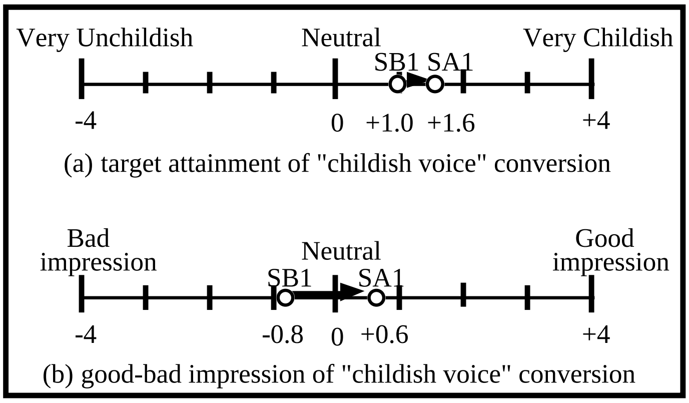


Figure 9: The results of the evaluation of target attainment and good-bad impression. The values averaged for all subjects are presented.

Next, we consider the results of the evaluation of target attainment and good/bad impression. The values averaged for all subjects are presented in Fig. 9. Relative to an attainment rate of +1.0 for the prosody modification coefficient combination obtained by a designer according to experience, the attainment rate for the evolutionary approach was 1.6, or an improvement of 0.6. For the impression evaluation, the scores were -0.8 for the human design approach and +0.6 for the evolutionary computation approach, or an improvement of 1.6. We believe that the reason for these results is that there was a strong tendency to raise the pitch in the adjustment by the designer to achieve the “childish voice” modification, resulting in a mechanical quality that produced an unnatural impression. The evolutionary computation approach, on the other hand, resulted in a modification that matched the objective without noticeable degradation in sound quality, and thus did not give the impression of processed voice.

4.2 EXPERIMENT WITH SYNTHESIZED SPEECH AS INPUT DATA

4.2.1 Voice Stimuli

The voice stimuli used in this experiment were as follows. Voice sample *S*1 consisted of the words “voice conversion using evolutionary computation of prosodic control” mechanically synthesized from text using Macintosh provided software (Macin Talk3). Voice samples *SC*1 to *SC*3 were obtained by performing voice conversion on the above sample for the three emotive terms of “childish,” “intelligible,” and “masculine” applying prosody modification coefficients obtained by the learning system using evolutionary computation as described above.

4.2.2 Experimental Method

As in the experiment using original speech, the subjects were 10 randomly selected males and females between the ages of 20 and 30 knowing nothing about the purpose of the experiment. Voice sample pairs *S*1 and *SCi* (*I*= 1-3) were presented through a speaker to these 10 subjects who were asked to judge whether voice conversion had succeeded in representing the above three emotive terms. This judgement was made in a three-level manner by selecting one of the following three responses: “close to the target expressed by the emotive term,” “can’t say,” and “very unlike the target.” Furthermore, for the sake of obtaining a quantitative comparison with respect to the emotive term “intelligible,” we also had the subjects perform a nine-level evaluation for both degree of attainment in voice conversion and good/bad impression for this voice sample pair. Subjects were allowed to hear each sample pair several times.

4.2.3 Experimental Results

The judgments of all subjects for voice sample pairs S1 and *SCi* (*i* = 1-3) are summarized in Fig. 10 in the form of a histogram for the responses “close to the target” and “very unlike the target.” These results demonstrate that voice conversion is effective for all emotive terms on average.

Figure 11 shows the results of judging degree of attainment and reporting good/bad impression averaged for all subjects. We see that degree of attainment improved by +1.2 from a value of +0.0 before conversion by determining an optimum combination of prosody modification coefficients using evolutionary computation. We also see that good/bad impression improved by +0.8 changing from +0.6 to +1.4.

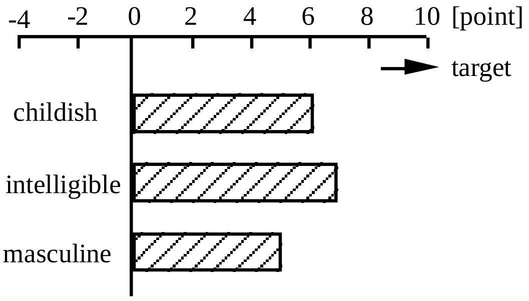


Figure 10: The results of the judgments of all subjects for voice sample pairs. The results are presented as a histogram for the responses "Close to the target" and "Very unlike the target".

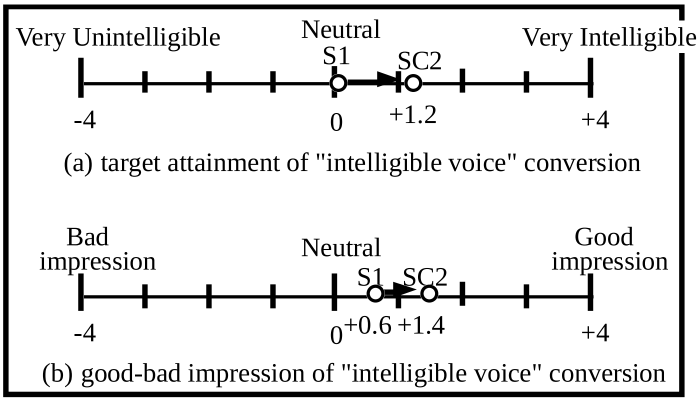

Very Unintelligible
Neutral
S1
SC2
Very Intelligible
-4
0
+1.2
+4
(a) target attainment of "intelligible voice" conversion
Bad impression
Neutral
Good impression
S1
SC2
-4
0
+0.6
+1.4
+4
(b) good-bad impression of "intelligible voice" conversion

Figure 11: The results of the evaluation of target attainment and good-bad impression. The values averaged for all subjects are presented.

5 DISCUSSION

The above experiments have shown that voice conversion using evolutionary computation can get closer to a target than parameter adjustment based on a designer's experience or trail and error. They have also shown that degradation in sound quality is relatively small and that listeners are not given a strong impression of a processed voice in the case of evolutionary computation. We here examine the question as to why evolutionary computation is superior. First, we consider the problem of accuracy in prosody modification coefficients. In the past, coefficients have been adjusted manually using real numbers of two or three significant digits such as 1.5 and 2.14. Such manual adjustment, however, becomes difficult if the search space becomes exceedingly large. On the other hand, it has been observed that a slight modification to a prosody modification coefficient can have a significant effect on voice conversion. For example, while raising pitch is an effective way of making a voice "childish," increasing the pitch modification factor gradually while keeping the amplitude modification factor and continuation time factor constant can suddenly produce an unnatural voice like that of a "spaceman." This can occur even by making a slight modification to the fourth or fifth decimal place. In other words, there are times when the accuracy demanded of prosody modification coefficients will exceed the range of manual adjustment.

Second, we consider the fact that each type of prosody information, that is, pitch, amplitude, and time continuation, is not independent but related to the other types. When manually adjusting coefficients, it is common to determine optimum coefficients one at a time, such as by first adjusting the pitch modification factor while keeping the amplitude modification factor and continuation time factor constant, and then adjusting the amplitude modification factor.

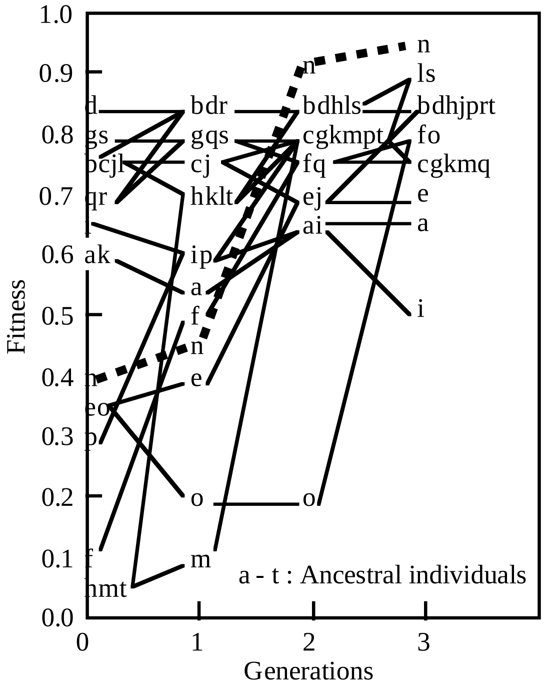


Figure 12: Fitness landscapes for "childish voice" conversion.

However, as pitch, amplitude, and time continuation are not independent of each other but exhibit correlation, it has been observed that changing the amplitude modification factor after setting an optimum value for the pitch modification factor will consequently change the optimum solution for pitch. This suggests that the modification coefficients for pitch, amplitude, and continuation time must be searched for in parallel.

Third, we consider the problem of multimodality accompanied by time fluctuation. For example, it often happens that a subject may not necessarily find an optimum solution from a voice that has already been subjected to several types of conversion. It has also been observed that optimum solutions may vary slightly according to the time that experiments are held and the physical condition of subjects at that time. In other words, we can view the problem as being one of determining a practical semi-optimum solution in as short a time as possible from a search space having multimodality and temporal fluctuation in the difficulty of prediction.

On the basis of the above discussion, we can see that the problems of voice conversion are indeed complex. For one, a practical semi-optimum solution must be determined in as short a time as possible from a search space having multimodality and temporal fluctuation in the difficulty of prediction. For another, high accuracy is demanded of modification coefficients and several types of modification coefficients must be searched for in parallel. In these experiments, we have shown that evolutionary computation is promising as an effective means of voice conversion compared to the complex real-world problems associated with finding an explicit algorithm and a solution based on trail and error by a designer. As a specific example, Fig. 12 shows the relationship between number of generations and fitness with respect to a “childish voice.” Ancestral individual information is shown from “a” to “t”. Here, individuals having prosody modification coefficients determined by experience are placed in the vicinity of a local optimum solution, and it takes only three generations to converge to a practical solution by performing genetic manipulation between these individuals and other individuals whose prosody modification coefficients are randomly set. Please see the example of voice conversion provided at http://webclub.kcom.ne.jp/ma/y-sato/demo/demo1.html for reference.

In future work, we will attempt to improve the accuracy of voice conversion by modifying spectral data as well, and must examine the application of evolutionary computation to parameter adjustment with the aim of synthesizing truly natural voices from arbitrary text. In this experiment, people evaluate results subjectively (based on feelings) and assign a fitness value to each individuals, it is also possible to select the individuals to be culled by a tournament system. It is also important to compare with other Evolutionary Computation method [Bäck 1997].

6 CONCLUSIONS

We have proposed the application of evolutionary computation to the adjustment of prosody modification coefficients for voice conversion, and have conducted voice-conversion experiments on both original speech recorded by a microphone and speech mechanically synthesized from text to evaluate the effectiveness of the proposed method. The results of these experiments revealed that adjustment of prosody modification coefficients by evolutionary computation performs voice conversion more efficiently than manual adjustment, and that degradation in sound quality is relatively small with no impression of a processed voice in the case of evolutionary computation. Future research must work on improving the accuracy of voice conversion by modifying spectral data as well, and must examine the application of evolutionary computation to parameter adjustment with the aim of synthesizing truly natural voices from arbitrary text.

Acknowledgments

The author would like to express their gratitude to Dr. Y. Kitahara, Mrs. H. Ando, and Mrs. M. Sato of Hitachi, Ltd., for their valuable discussions on experimental results.

References

T. Bäck, U. Hammel and H.-P. Schwefel (1997). Evolutionary Computation: Comments on the History and Current State, *IEEE Trans. on Evolutionary Computation*, Vol.1, No.1, pp.3-17.

D.E. Goldberg (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley.

Y. Kitahara and Y. Tohkura (1992). Prosodic Control to Express Emotions for Man-Machine Speech Interaction”, *IEICE Trans. Fundamentals.*, Vol. E75, No. 2, pp. 155-163.

D.H. Klatt and L.C. Klatt (1990). Analysis, Synthesis, and Perception of Voice Quality Variations among Female and Male Talkers, *Jounal of Acoustic Society America*, 8

J D. Malah (1979). Time-domain algorithms for harmonic bandwidth reduction and time scaling of speech signals”, *IEEE Trans. Acoust. Speech, Signal Processing*, Vol. ASSP-27, pp. 121-133.

Y. Sato (1997). Voice Conversion Using Evolutionary Computation of Prosodic Control, in *Proc. of the Australasia-Pacific Forum on Intelligent Processing and Manufacturing of Materials*, pp. 342-348.

K. Sims (1991). Interactive Evolution of Dynamical Systems, in F.J. Varela and P. Bourgine (eds.), Toward a Practice of Autonomous Systems, in *Proc. of the First European Conf. on Artificial Life*, MIT Press, pp.171-178.

H. Takagi (2001): Interactive Evolutionary Computation: Fusion of the Capabilities of EC Optimization and Human Evaluation, *Tutorial Book of the 2001 Congress on Evolutionary Computation.*