EC Application in Speech Processing - Voice Quality Conversion Using Interactive Evolution of Prosodic Control

Yuji Sato
Hosei University
Japan

1. Introduction

This chapter outlines Interactive Evolutionary Computation (IEC) and IEC application in speech processing. IEC is an optimization method that adapts evolutionary computation (EC) based on subjectively human evaluation. EC is a biologically inspired general computational algorithm and includes genetic algorithms (GA), genetic programming (GP), evolution strategies (ES), and evolutionary programming (EP) (Bäck et al., 1997). These algorithms are a heuristic search technique based on the ideas of natural selection and work with a population of solutions that undergo modifications under the influence of genetic manipulations and finally converge to the optimal solution. Here, the selection will be processed based on an explicit evaluative function prepared by human designer beforehand. EC has been widely used to engineering applications and effective for a variety of complex large sized combinational problems such as the travelling salesman problem, the job-shop scheduling problem, the machine-loading problem, etc. On the other hand, it is difficult, or even impossible, to design human evaluation explicit functions for interactive systems. For example, to obtain the most favorable outputs from interactive systems that create or retrieve graphics or music, such outputs must be subjectively evaluated. IEC is the technology that EC optimizes the target systems based on subjective human evaluation as fitness values for system outputs and effective for such kind of interactive systems. IEC research conducted during 1990s originated from the work of Dawkins (Dawkins, 1986) and two major research streams developed during the 1990s. The first research stream (Sims, 1991; Biles, 1994; Unemi, 2002) is Artificial Life (Dawkins, 1989) and the second research stream comes from the increase of researchers who are interested in humanized technology or human-related systems and have applied the IEC to engineering fields (Watanabe & Takagi, 1995; Takagi & Ohya, 1996; Sato, 1997; Parmee & Bonham, 1999; Kim & Cho, 2000). In the following, as a concrete example, we describe about voice quality conversion using IEC.

In section 2 we briefly explain about a basic voice conversion technique and application. In section 3 we show voice elements and our voice quality conversion systems. In section 4 we describe prosodic coefficient fitting by IEC. In section 5 we describe the simulations we performed to evaluate this technique, and in the final sections we discuss our results and finish with a conclusion.

Source: Evolutionary Computation, Book edited by: Wellington Pinheiro dos Santos,
ISBN 978-953-307-008-7, pp. 572, October 2009, I-Tech, Vienna, Austria

2. Voice conversion techniques and applications

With the sudden arrival of the multimedia era, the market for multimedia information devices centered about the personal computer has been growing rapidly. The market for multimedia application software has likewise taken off providing an environment where users can manipulate images and sound with ease. At the same time, new markets are appearing using voice conversion (processing) technology.

Voice conversion is a technique for transforming the timbre and other characteristics of voices. Research into voice conversion has so far been centered on techniques for transforming the voice of one speaker (A) into the voice of another (B), principally with the aim of making synthetic speech sound more natural by giving it some sort of personality. For example, since personality parameters are difficult to specify, voice conversion between designated speakers A and B is performed by associating the voice patterns with VQ codebooks for each speaker (Shikano et al., 1991; Moulines & Sagisaki, 1995; Levent & David, 1997). These VQ codebooks are made by quantizing the feature space of the speaker's vocal characteristics as a set of vectors which are then encoded. Research is also being carried out into voice morphing techniques (Slaney et al., 1996) that can achieve a continuous transformation from speaker A to speaker B. Voice morphing is implemented by continuously varying the correspondence between speakers A and B with regard to some attribute such as a spectral envelope. Techniques associated with voice conversion are summarized in the reference by Puterbaugh (Puterbaugh, 2009).

However, these voice conversion and voice morphing techniques often have residual problems with the resulting speech quality due to the large transformation that results from converting speech characteristics down to the parameter level. Effort must also be put into preliminary tasks such as acquiring speech data from the target speaker of the conversion, and having a speech specialist generate a VQ codebook from this data. If a technique can be found to enable ordinary users to convert freely between voices of their own choosing without any particularly noticeable audio degradiation and without being limited to conversion between specific pre-registered speakers or having to prepare resources such as a VQ codebook, then a much wider range of marketable applications would be opened up to voice conversion techniques.

These include multimedia-content editing, computer games, and man-personal machine interfaces. 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 conversion can be felt here. Technology for converting raw speech to clear and easily understand speech can also be effective for people that are required to give public speeches. In this regard, we can cite the example of Margaret Thatcher, the former U.K. prime minister, who took voice training so that she could give speeches in a clear, resonant voice that could provide more impact (Sample, 2002). Voice conversion technology can also be useful in the world of computer games, especially for multiplayer configurations. For example, a player puts on a headphone set and speaks into a microphone, and the input speech is conveyed to other players in the game as the voice of the game character representing the incarnation of that player. Voice conversion technology could be used here to convert raw input speech to an extremely clear voice rich in intonation and emotion making for a more compelling game. Finally, in the world of man-personal machine interfaces, voice-based interfaces have

been considered for some time, and big markets are now anticipated for a wide range of voice-based applications from the reading out of e-mail and World Wide Web text information to the voice output of traffic reports and other text information by car-navigation equipment. It is essentially impossible, though, to set a voice favorable to all users beforehand on the manufacturer's side, and if technology existed that could enable the user to convert speech to a voice of his or her liking, it might have a great effect on such applications.

There are still some issues, however, that must be overcome in products using the voice processing technologies described above. To begin with, only qualitative know-how has so far been acquired with respect to voice processing technologies that convert raw speech to a clear voice or a voice that one desires, and specific parameter setting is performed on a trial and error basis. Optimal parameter values for voice conversion are also speaker dependent, which makes parameter adjustment difficult. Nevertheless, in voice output equipment targeting the readout of e-mail and World Wide Web text information, products using speech synthesis technology are beginning to be developed. Mechanically synthesized speech using rule-based synthesis, however, suffers from various problems, including an impression of discontinuity between phoneme fragments, degraded sound quality due to repeated signal processing, and limitations in sound-source/articulation segregation models. In short, speech synthesis that can produce a natural sounding voice is extremely difficult to achieve. Current speech-synthesis technology tends to produce mechanical or even unintelligible speech, and voice-quality problems such as these are simply delaying the spread of speech synthesis products.

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 interactive evolutionary computation to parameter adjustment for the sake of voice quality 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). 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 quality conversion but also of synthetic speech generated automatically from text data (Sato, 2002). This chapter shows those experimental results and discuss why parameter adjustment using evolutionary computation is more effective than that based on trial and error by an experienced designer.

3. Voice elements and their relationship to voice quality conversion

3.1 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 & Klatt, 1990). In other words, we can consider prosodic information and spectral information as feature quantities for the control of voice quality. Prosody consists of three main elements—pitch information, amplitude information, and temporal structure—described as follows. First, pitch is the basic frequency of the speech waveform, that is, the frequency at which the vocal chords vibrate to generate sound, and it carries information related to voice modulation. The accent in a word and intonation in a sentence are formed by fluctuation in pitch information. In human speech, average pitch is about 130 Hz in men and 250 Hz in women. Next, amplitude information refers to the actual amplitude of the speech waveform, that is, the vibrational energy of the vocal chords, and it carries information related to vocal strength. Amplitude information is used to control stress placed on a word and emphasis placed at various levels when speaking. While there are several methods for expressing amplitude information, we will here use the "short-interval average power" method. Finally, temporal structure corresponds to the temporal operation of the vocal chords, and this structure is responsible for forming rhythm in speech and generating pauses, for example. Temporal structure is thought to contribute to detection of important words.

3.2 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 & Tohkura, 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 quality conversion by limiting the data to be controlled to pitch data, amplitude data, and temporal structure prosodic data.

Figure 1 shows the pitch modification method. Pitch modification is not performed by modifying temporal length. Rather, when raising pitch, for example, the procedure is to repeat the waveform partially cut from one pitch unit and then insert the same waveform as that of the prior interval every so many cycles of the above process. This does not change temporal length. Then, when lowering pitch, the procedure is to insert a mute portion into each pitch unit and then curtail the waveform every so many cycles of the above process so that again temporal length does not change. Next, Fig. 2 shows the method for converting temporal length. In this method, temporal length is converted by extension or contraction without changing pitch by the time-domain-harmonic-scaling (TDHS) (Malah, 1979) enhancement method. In Fig. 2(a), where one pitch period (autocorrelation period) is denoted as Tp and the extension rate as γ, shift Ls corresponding to the extension rate can be expressed by Eq. (1) below.

$$Ls = \frac{Tp}{\gamma - 1} \tag{1}$$

Likewise, in Fig. 2(b), where the contraction rate is denoted as γ, shift Lc corresponding to the contraction rate can be expressed by Eq. (2).

$$Lc = \frac{\gamma Tp}{\gamma - 1} \tag{2}$$

Amplitude is controlled by converting on a logarithmic power scale. Letting W_i denote the current value and β the modification coefficient, the modification formula is given by Eq. (3) below.

$$\log_{10} W_{i+1}^2 = \log_{10} W_i^2 + \beta \tag{3}$$

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 by learning. As the learning algorithm, this unit employs evolutionary computation.

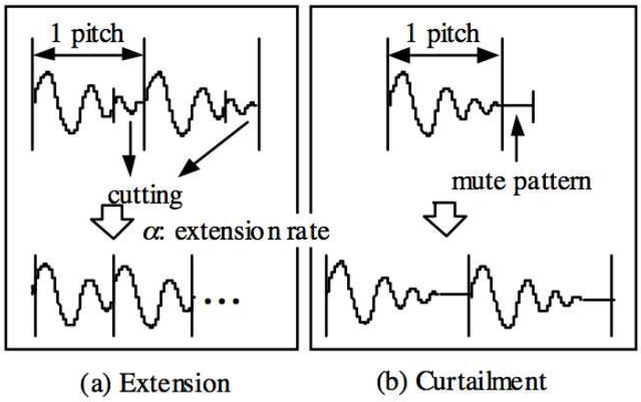

Fig. 1. Extension and curtailment of pitch period. Pitch is raised by cutting out of the waveform within one pitch unit. Pitch is lowered by inserting silence into pitch unit.

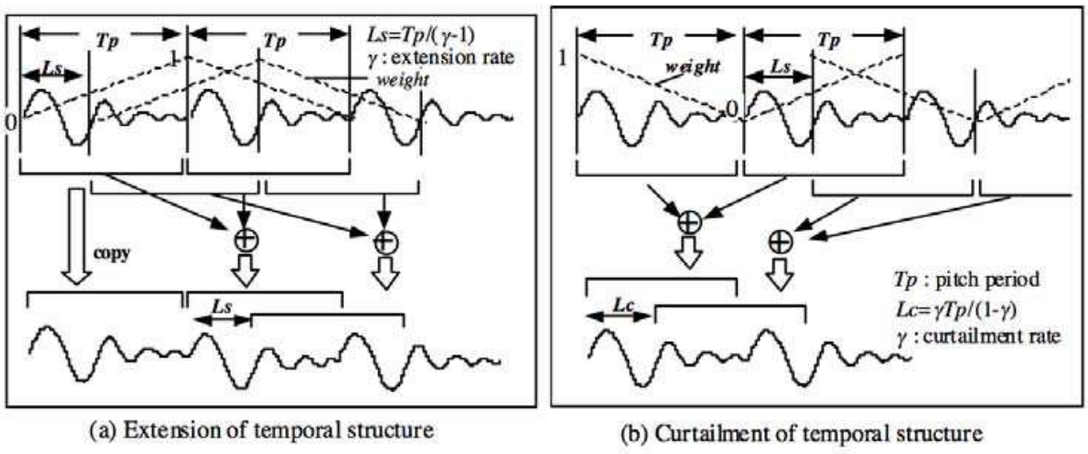

Fig. 2. Extension and curtailment of temporal structure. The continuation length is accomplished by using the TDHS enhancement method to extend or contract the sound length without changing the pitch.

3.3 Formulation of the voice quality conversion problem

First, the objective function f is unknown, and the optimal solutions of α, β and γ are speaker-dependent. For example, the algorithm for specifically determining the value of α is unknown, and the optimal value of α changes depending on the speaker of the speech waveform to be transformed. Second, the optimal values of variables α, β and γ are not entirely independent, and are weakly correlated to one another. For example, experiments have shown that when an attempt is made to find an optimal solution for α with fixed values of β and γ, this optimal solution for α will change slightly when the value of β is subsequently altered. Third, since the evaluation is performed at the level of the objective function f instead of the variables α, β and γ, the solution of this problem involves the use of implicit functions. At last, 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.

The voice quality conversion problem is therefore formulated as follows:

$$\left.\begin{array}{rc} \textit{Minimize} & f(\alpha,\beta,\gamma,t) \\ \textit{subject to} & \alpha = g_1(\beta,\gamma),\ \beta = g_2(\gamma,\alpha),\ \gamma = g_3(\alpha,\beta) \\ & (\alpha,\beta,\gamma) \in X = R^n \end{array}\right\} \tag{4}$$

where α, β and γ are conversion coefficients for pitch, power and time duration, respectively. Here, $X = R^n$ is an n-dimensional real space, and $f(\alpha, \beta, \gamma, t)$ is a real-value function of multimodality accompanied by time fluctuation. And f, g_1, g_2, g_3 are unknown (and probably non-linear) functions.

4. Prosodic coefficient fitting by IEC

4.1 Configuration of the voice quality conversion system

The configuration of the voice quality conversion 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.

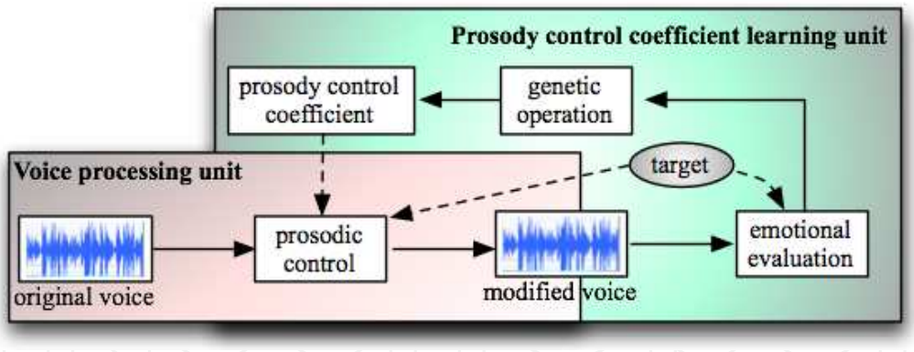

Fig. 3. Block diagram of proposed voice quality conversion system. The system comprises a voice processing part and prosody control coefficient learning part.

4.2 Overview of interactive evolution of prosodic control

The first step in this IEC procedure is to define chromosomes, i.e., to substitute the search problem for one of determining an optimum chromosome. In this procedure, we define a chromosome as a one-dimensional real-number array corresponding to a voice-modification 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 (Takagi, 2001) in which people evaluate results subjectively (based on feelings) for each generation.

4.3 Genetic manipulation

4.3.1 Selection rule

The population is replenished by replacing the culled individuals with a new generation of individuals picked by roulette selection (Holland, 1975; Goldberg, 1989) based on human evaluation.

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.

4.3.2 Crossover and mutation

Figure 4 presents an example of the proposed crossover and mutation operation. In the crossover operation, any one column is chosen and the values in that column are swapped in the two parent individuals. Here, 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 this operation, one of the two entities generated by the crossover operation is randomly subjected to crossovers in which the average value of each coefficient in the two parent entities (Bäck, 2000) are obtained.

For spontaneous mutations, the standard deviation of the mutation distribution was set small as shown in Eq. (5) for entities where crossovers were performed just by swapping coefficients as in the conventional approach. In this equation, C_i represents a modification coefficient for generation i, I is a unit matrix, and N is a normal distribution function with a mean vector of 0 and a covariance of 0.000025I.

$$C_{i+1} = C_i + N(0, 0.000025I) \tag{5}$$

Conversely, a larger standard deviation was set for entities where crossovers were performed by taking the average of two coefficients as shown in Eq. (6).

$$C_{i+1} = C_i + N(0, 0.01I) \tag{6}$$

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.

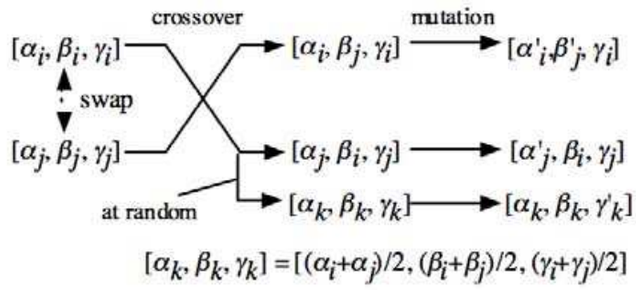

Fig. 4. Example of the proposed genetic manipulation. In this operation, one of the two entities generated by the swap operation is randomly subjected to crossovers in which the average value of each coefficient in the two parent entities are obtained.

5. Evaluation experiments

5.1 Experiment with original speech as input data

5.1.1 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.

5.1.2 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 $S0$ together with SA_i (i = 1 to 5) and $S0$ together with SB_i (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.

5.1.3 Experimental results

The results of the judgments of all subjects for voice sample pairs $S0$ - SA_i (i = 1 to 5) and $S0$ - SB_i (i = 1 to 5) are presented in Fig. 5 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.

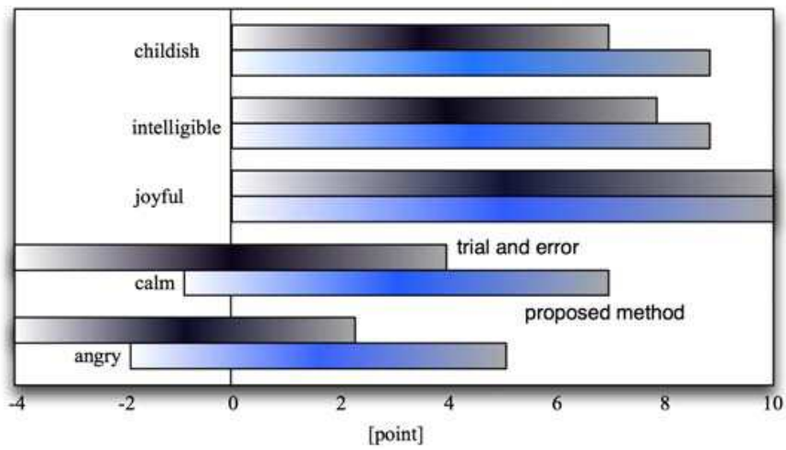

Fig. 5. The results of the judgment of all subjects for voice sample pairs. The results are presented as a histogram for the responses "Close to the target" and "Very unlike the target"

Next, this section shows the results of the evaluation of target attainment and good/bad impression. The values averaged for all subjects are presented in Fig. 6. Relative to an attainment rate of +1.2 for the prosody modification coefficient combination obtained by a designer according to experience, the attainment rate for the evolutionary approach was 1.8, or an improvement of 0.6. For the impression evaluation, the scores were -0.6 for the human design approach and +0.8 for the evolutionary computation approach, or an improvement of 1.4. 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.

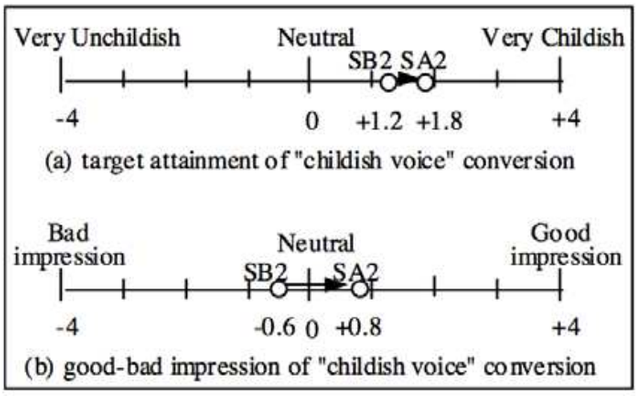

Fig. 6. The results of the evaluation of target attainment and good-bad impression. The values averaged for all subjects are presented.

5.2 Experiment with synthesized speech as input data

5.2.1 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.

5.2.2 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 SC_i (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.

5.2.3 Experimental results

The judgments of all subjects for voice sample pairs *S*1 and SC_i (i = 1-3) are summarized in Fig. 7 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 8 shows the results of judging degree of attainment and reporting good/bad impression averaged for all subjects. These results demonstrate 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.

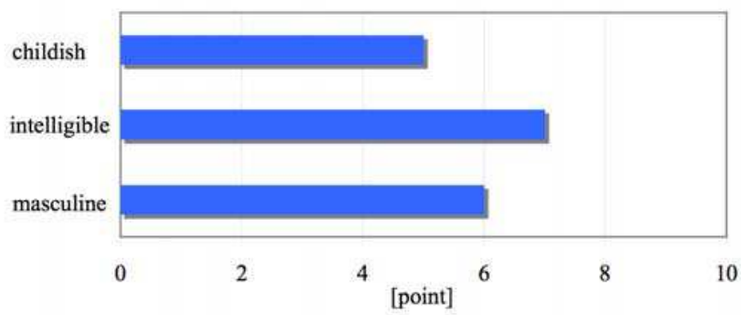

Fig. 7. The results of the judgment of all subjects for voice sample pairs. The results are presented as a histogram for the responses "Close to the target" and "Very unlike the target"

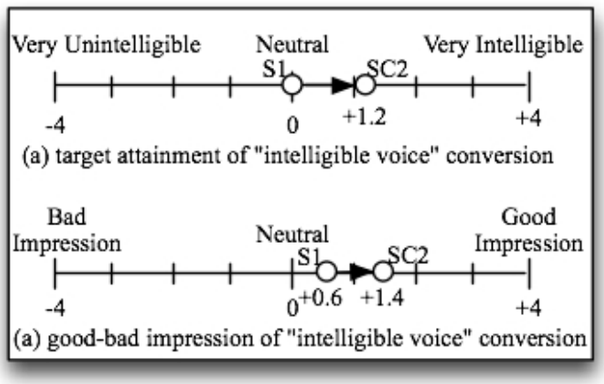

Fig. 8. The results of the evaluation of target attainment and good-bad impression. The values averaged for all subjects are presented.

6. Discussion

The above experiments have shown that voice quality 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 quality 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. 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 quality 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 quality 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.

Because the current system employs an interactive type of evolution, a real-time search for optimal modification coefficients for prosodic control cannot be performed. The system also suffers from various limitations, such as input data taken to be the utterance of one sentence within a fixed period of time. On the other hand, when conducting an experiment to check for sentence dependency of optimal modification coefficients when targeting the emotive terms of “intelligible,” “childish,” and “masculine” no particular sentence dependency was found for the same speaker. In other words, optimal modification coefficients for voice quality conversion, while being speaker dependent, are considered to be sentence independent, or any text dependency that exists is within a range that can be ignored for the most part. Thus, once optimal modification coefficients for converting raw speech to intelligible speech can be found for a certain sentence, the values of those coefficients can be applied to other sentences as long as the speaker remains the same. In addition, voice quality conversion after optimal modification coefficients have been determined can be executed in real time from stored input data, which points to the feasibility of applying current technology to real products in fields like addition of narration in multimedia-content editing, computer games, recorded TV soundtracks, and man-personal machine interfaces given that appropriate procedures for use are established. Please refer to http://www.h3.dion.ne.jp/~y-sato/demo/demo1.html for an example of voice quality conversion.

7. Conclusion

We have proposed the application of interactive evolutionary computation to the adjustment of prosodic modification coefficients for the purpose of voice quality conversion. To examine the effectiveness of the technique described, we performed voice quality conversion experiments on both original human speech recorded by a microphone and speech mechanically synthesized from text, and evaluated results from the viewpoint of target attainment and good or bad impression. It was found that the application of interactive evolutionary computation could convert speech more efficiently than manual adjustment of prosodic modification coefficients. Furthermore, on comparing speech obtained by the proposed technique with that obtained by manual determination of prosodic modification coefficients, it was found that the former exhibited relatively little deterioration in voice quality and no impression of processed speech. Given, moreover, that the values of optimal parameters, while speaker dependent, are sentence independent, and that voice quality conversion can be executed online from stored input data once optimal parameters have been determined, we expect this technique to be applicable to products in fields such as addition of narration in multimedia-content editing, computer games, and man-personal machine interfaces even with current technology. Future studies must take up means of improving the accuracy of voice quality conversion by adding the conversion of spectral information and the application of evolutionary computation to parameter adjustment for synthesizing natural speech from continuously input text.

8. Acknowledgments

The author would like to express their gratitude to Dr. Y. Kitahara and Mrs. H. Ando of Hitachi, Ltd., Central Research Laboratory, and Dr. M. Sato of Tokyo University of Agriculture & Technology for their valuable discussions on experimental results.

9. References

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

Bäck, T.; Fogel, D.B. & Michalewicz, Z. (2000). *Evolutionary Computation 1: Basic Algorithms and Operators*, 2000, Institute of Physics Publishing, Bristol, UK.

Biles, J.A. (1994). GenJam: A Genetic Algorithm For Generating Jazz Solos, In: *Proceedings of the 1994 International Computer Music Conference* (ICMC-94), pp.131-137, Aarhus, Denmark, 1994.

Dawkins, R. (1986). *The Blind Watchmaker*, Long-man, Essex, 1986.

Dawkins, R. (1989). The Evolution of Evolvability, Langton, C.G. (ed.), *Artificial Life*, pp.201-220, 1989, Addison-Wesley, Reading, MA.

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

Holland, J.H. (1975). *Adaptation in Natural and Artificial Systems*, 1975, University of Michigan Press (Second edition: 1992, MIT Press, Cambridge, MA).

Kim, H.-S. & Cho, S.-B. (2000). Knowledge-based encoding in interactive genetic algorithm for a fashion design aid system, In: *Proceedings of the 2000 Genetic and Evolutionary Computation Conference*, p. 757, Las Vegas, July 2000, Morgan Kaufmann Publishers.

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

Klatt, D.H. & Klatt, L.C. (1990). Analysis, Synthesis, and Perception of Voice Quality Variations Among Female and Male Talkers, *Journal of Acoustic Society America*, 87(2), pp. 820-856, 1990.

Koza, J.R. (1992). *Genetic Programming: On the Programming of Computers by Means of Natural Selection*, 1992, MIT Press, Cambridge, MA.

Levent, M.A. & David, T. (1997). Voice Conversion by Codebook Mapping on Line Spectral Frequencies and Excitation Spectrum, In: *Proceedings of the EuroSpeech97*, 1997.

Malah, J.D. (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, 1979.

Moulines, E. & Sagisaka, Y. (1995). Voice Conversion: State of Art and Perspectives, *Speech Communication 16*, pp. 125-126, 1995.

Parmee, I.C. & Bonham, C.R. (1999). Cluster-oriented genetic algorithms to support interactive designer/evolutionary computing systems, In: *Proceedings of the Congress on Evolutionary Computation*, pp. 546-553, 1999.

Puterbaugh, J. (2009). Dr. John Puterbaugh homepage at the Department of Music, Princeton University. <http://silvertone.princeton.edu/~john/voiceconversion.htm> Accessed on: June 9, 2009.

Sample, I. (2002). Darwinian boost for public speakers, *NewScientist*, Vol. 175, No. 2352, p. 18, 2002, New Science Publications, London.
(http://www.newscientist.com/news/news.jsp?id=ns99992560)

Sato, Y. (1997). Voice Conversion Using Evolutionary Computation of Prosodic Control, In: *Proceedings of the Australasia-Pacific Forum on Intelligent Processing and Manufacturing of Materials* (IPMM-97), pp. 342-348, Gold Coast, Australia, July 1997.

Sato, Y. (2002). Voice Conversion Using Interactive Evolution of Prosodic Control, In: *Proceedings of the 2002 Genetic and Evolutionary Computation Conference* (GECCO-2002), pp. 1204-1211, New York, July 2002, Morgan Kaufmann Publishers, San Francisco, CA.

Shikano, K.; Nakamura, S. & Abe, M. (1991). Speaker Adaptation and Voice Conversion by Codebook Mapping, In: *Proceedings of the IEEE Symposium on Circuits and Systems*, Vol. 1, pp. 594-597, 1991.

Sims, K. (1991). Interactive Evolution of Dynamical Systems, Varela, F.J. & Bourgine, P. (eds.), Toward a Practice of Autonomous Systems, In: *Proceedings of the First European Conference on Artificial Life*, pp.171-178, 1991, MIT Press, Cambridge, MA.

Slaney, M.; Covell, M. & Lassiter, B. (1996). Automatic Audio Morphing, In: *Proceedings of the ICASSP*, pp. 1001-1004, 1996.

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

Takagi, H. & Ohya, K. (1996). Discrete Fitness Values for Improving the Human Interface in an Interactive GA, In: *Proceedings of the IEEE 3rd Inter. Conf. on Evolutionary Computation*, pp. 109-112, Nagoya, May 1996, IEEE Press.

Unemi, T. (2002). SBEAT3: A Tool for Multi-part Music Composition by Simulated Breeding, In: *Proceedings of the A-LIFE VIII*, pp. 410-413, Sydney, NSW, Australia, 2002.

Watanabe, T. & Takagi, H. (1995). Recovering System of the Distorted Speech Using Interactive Genetic Algorithms, In: *Proceedings of the IEEE Inter. Conf. on Systems, Man and Cybernetics*, Vol. 1, pp. 684-689, 1995.

Evolutionary Computation

Edited by Wellington Pinheiro dos Santos

ISBN 978-953-307-008-7
Hard cover, 572 pages
Publisher InTech
Published online 01, October, 2009
Published in print edition October, 2009

This book presents several recent advances on Evolutionary Computation, specially evolution-based optimization methods and hybrid algorithms for several applications, from optimization and learning to pattern recognition and bioinformatics. This book also presents new algorithms based on several analogies and metafores, where one of them is based on philosophy, specifically on the philosophy of praxis and dialectics. In this book it is also presented interesting applications on bioinformatics, specially the use of particle swarms to discover gene expression patterns in DNA microarrays. Therefore, this book features representative work on the field of evolutionary computation and applied sciences. The intended audience is graduate, undergraduate, researchers, and anyone who wishes to become familiar with the latest research work on this field.

How to reference

In order to correctly reference this scholarly work, feel free to copy and paste the following:

Yuji Sato (2009). EC Application in Speech Processing - Voice Quality Conversion Using Interactive Evolution of Prosodic Control, Evolutionary Computation, Wellington Pinheiro dos Santos (Ed.), ISBN: 978-953-307-008-7, InTech, Available from: http://www.intechopen.com/books/evolutionary-computation/ec-application-in-speech-processing-voice-quality-conversion-using-interactive-evolution-of-prosodic

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen