

Recovering System of the Distorted Speech using Interactive Genetic Algorithms

Tatsumi WATANABE
Matsushita Electric Industrial Co., Ltd.
3-4, Hikaridai, Seika, Souraku,
Kyoto, 619-02 Japan
wata@crl.mei.co.jp

Hideyuki TAKAGI
Kyushu Institute of Design,
9-1, Shiobara 4-chome, Minami-ku,
Fukuoka, 815 Japan
takagi@artemis.kyushu-id.ac.jp

Abstract

This paper investigates the application of an interactive GA to speech signal processing. Three users design a filter to enhance speech using the interactive GA. Quality of the three 3 filters are evaluated on speech samples by 32 subjects on the basis of 2 subjective criteria. Our results show that the interactive GA is effective for this task from a statistical test point of view.

1 Introduction

There are many optimization tasks which can not use implicit cost or objective functions: for example, designing interior or industrial equipment under a given concept, or tuning systems that output acoustic or image according to human preference. However, we cannot use conventional optimization methods for these tasks which require deep analytical information, such as derivative information, continuity condition, or complete knowledge of the tasks. Only humans can efficiently evaluate these tasks that involve human perception. If the mathematical fitness function can be replaced by a human, these tasks can be optimized systematically.

Interactive genetic algorithms (interactive GA) have been proposed as GA [1, 2] that uses a human evaluation as the fitness function. We apply interactive GA to a problem in the class of tasks mentioned above.

Research that use the interactive GA have been presented. Examples of interactive GA research are: montage face image generation[3], designing shapes of bug biomorphs[4], creating images, evolving expressions that specify particular sequences of image-processing functions[5], combining interactive evolution with constructive solid geometry techniques to create computer renderings of three dimensional forms ("virtual sculptures")[6], general-purpose interactive graphic layout system based on GA[7], the design of a double curvature concrete arch dam[8], line drawing and application to face drawing[9], and the decision supporting

⁰This paper will appear in the Proceeding of IEEE Int'l Conf. on System, Man & Cybernetics, Vancouver, Canada, Oct. 22-25, 1995.

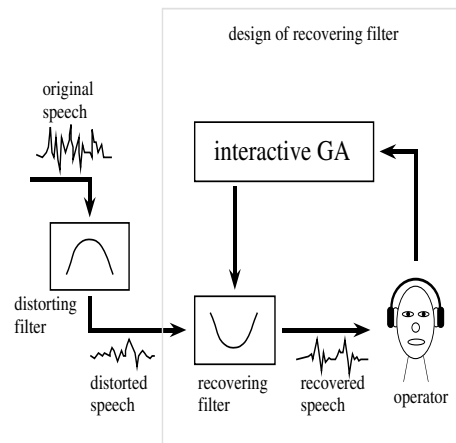


Figure 1: The system diagram of recovering distorted speech using interactive GA.

system for aesthetic design of cable-stayed bridges[10].

Most of the research applied the interactive GA to dealing with 2-dimensional images and figures which can be simultaneously and spatially shown to a human operator. There were few tasks where the interactive GA were applied to dynamic processes such as signal processing.

Our objective is to investigate whether the interactive GA is effective to use for signal processing. We apply it to the task of designing a filter that enhances distorted speech and study the effectiveness of our system using two subjective tests.

2 Experiment of recovering the distorted speech

2.1 Experimental system

Three users A, B and C design a recovering filter for distorted speech using interactive GA according to their perceptive evaluation hearing. Figure 1 shows total system used in our experiment.

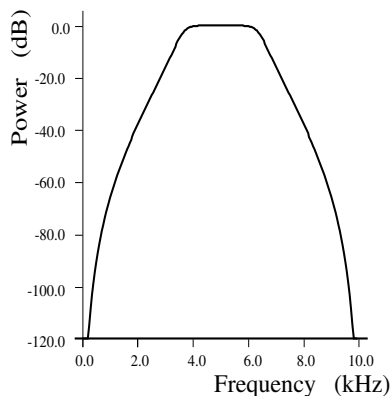


Figure 2: Characteristics of the distorting filter in this experiment.

Distorted speech

The 20kHz sampled Male speech “Doómo arigatoh gozaimashita” is used as original non-distorted speech. This speech means “thank you very much”. The length of speech is about 2.5 seconds. A distorting filter is designed using IIR (Infinite Impulse Response) filter to suppress the energy in lower and higher frequency of the original speech as shown in Figure 2.

Design of recovering filter

The distorted speech made by the distorting filter is inputted to a recovering filter to enhance the quality of speech. The design of the recovering filter is the task of the interactive GA in this experiment. Operators hear the recovered speech and evaluate its quality. The interactive GA tunes the parameters of the recovering filter according to operator’s evaluation. In each iteration of the interactive GA, it designs multiple recovering filters, operator hears and evaluates the recovered speeches, and the interactive GA makes the next generation of parameters of filters. This system repeats the above procedure until operator is satisfied with the quality of the recovered speech.

Design of gene coding

A chromosome includes eight genes which correspond to coefficients of FIR (Finite Impulse Response) filter. The transfer function of the filter is:

$$H(z) = \sum_{k=0}^7 a_k z^{-k}.$$

$\mathbf{a} = (a_0, a_1, \dots, a_7)$ is a parameter vector of the FIR filter. We use binary coding, and each a_k is coded by eight bits. The number of individuals per generation (Population size) is 20. The Interactive GA designs the filter coefficients \mathbf{a} according to human evaluation.

2.2 Combination of speech with image

One of difficulties in applying the interactive GA to problems involving dynamic tasks is that it is difficult for a user to compare several individuals simultaneously. In addition it can be difficult to remember the temporal features of previously rendered signals. The following two ideas are proposed to increase the performance of the interactive GA by assisting the operator’s memory and by reducing the experimental load of the operator.

As visual-aid for the speech signals, 20 line-drawn faces are combined with speech samples that 20 recovering filters generate and displayed spatially. The corresponding speech sample is played to an operator when he or she points a face. The size, angle, and position of eyebrows, eyes, nose, and mouth determine the impression of each face. 8 of these 18 parameters in a face correspond to 8 coefficients of a recovering filter. The other face parameters are fixed. The emotional impression of the face itself has no correlation to the quality of the corresponding filter and the recovered speech. However, the combination of sound and face images is expected to help operators to remember the impression of the speech sample to compare 20 recovering filters.

The second idea is to prepare the function of sorting the individual speech samples according to the user’s evaluation measure. Since the perceptual fitness measure may fluctuate, the 20 absolute fitness values given by the user may contain noise. Sorting the 20 speech samples allows the user to verify the relative ordering of fitness values is correct.

2.3 Operators of experimental system

Our interactive GA users are two males, A and B, and one female, C. All are between 20 and 30 years old. C is a student majoring in acoustics, while A and B are not.

The three users conduct the task of designing a recovering filter using interactive GA independently. The filter design process using our interactive GA took A and B 1.5 hours or 8 generations a day, while C took 2.5 hours or 8 generations a day. Each user rested as needed during the experiment. Each completed the task in five days evolving 40 generations.

Table 1: Rating category of the method of successive categories.

Rating category	score
Sonority is bad	-2
Sonority is a little bad	-1
Sonority is plain	0
Sonority is a little good	1
Sonority is good	2

2.4 Characteristics of the recovered speeches by human operators

We evaluate the recovery filters designed by the three interactive GA users. Figure 3 shows the long term power spectrum of the recovered speech made by A, B and C. Distorted speech, recovered speeches in 10th, 20th, and 40th generation are shown from the bottom. Each line in figure 3 is shifted and represented to facilitate comparisons.

Figure 3 shows that the recovering filters were designed to boost suppressed energy in low frequency band of the distorted speech (see inside the dotted ellipses in figure 3). On the contrary, the high frequency is boosted very little or even suppressed. These results are consistent with the well-known fact that low frequencies of human voice, where low order of formants exist, are much more important to human perception than high frequency[11].

3 Evaluation of recovered speeches

3.1 Subject test

To evaluate the quality of the recovery filters two subjective tests by 32 people were conducted: the method of successive categories [12] and the variation of Sheffé’s method of paired comparisons [13, 14].

Method of successive categories

In the method of successive categories, sonority of each speech is evaluated according to Table 1.

Sheffé’s method of paired comparisons

Sheffé’s method of paired comparisons is a subjective evaluation method that successively gives two speech samples, S_i and S_j , to subjects and requests to evaluate the preference of relative sonority of the two according to Table 2.

Table 2: Rating category of the Sheffé’s method of paired comparisons.

Rating category	score
S_j is worse than S_i	-2
S_j is a little worse than S_i	-1
S_j is almost equal to S_i	0
S_j is a little better than S_i	1
S_j is better than S_i	2

3.2 Subject tests condition

We conduct the subject tests with concerning the following 13 speech samples: (1) distorted speech generated by a distorting filter, (2-4) recovered speech samples of 3 subjects in 10th generation, (5-7) those in 20th generation, (8-10) those in 40th generation, and (11-13) 3 recovered speech samples generated by 3 $filter_{WLR}$ s, where $filter_{WLR}$ is a recovering filter designed by a conventional GA that has a fitness function of WLR (weighted likelihood ratio) distance [15] between recovered speech and original clear speech. (11) to (13) are not generated using the interactive GA, and are generated by giving perfect supervised speech, original speech. We discuss (11) to (13) in detail in section 4.

The procedure of the method of successive categories is the followings. 32 subjects evaluate the quality of 13 sets of speech sequentially according to Table 1. One set consists of the number in voice and the followed evaluation task of speech which is about 2.5 seconds. 13 sets are sequentially given to subjects with the interval of 3.0 seconds. This iteration is repeated three times. Subjects evaluate each by each and are requested to write the final evaluation for 13 speeches.

The procedure of Sheffé’s method is the followings. Two speech samples are selected among the distorted speech of (1) and recovered speech samples of (2-10) that same user generated; two are selected from four speech samples. They are combined in consideration of presenting order. These combinations are made for the three users. The total number of evaluation sets is: ${}_4C_2 \times 2 \text{ orders} \times 3 \text{ operators} = 36$. The 36 speech pair sets are presented to 32 subjects. One set consists of identification of the pair, the first speech sample of S_i , 0.4 seconds silence, and then the second speech sample of S_j . Each set is repeated at 1.6 second intervals. After the set is repeated three times, subjects are requested to evaluate the relation of S_i to S_j according to Table 2.

The test data obtained in the above ways are analyzed using two subjective tests. 32 subjects are divided evenly into four groups. Each group was seated about

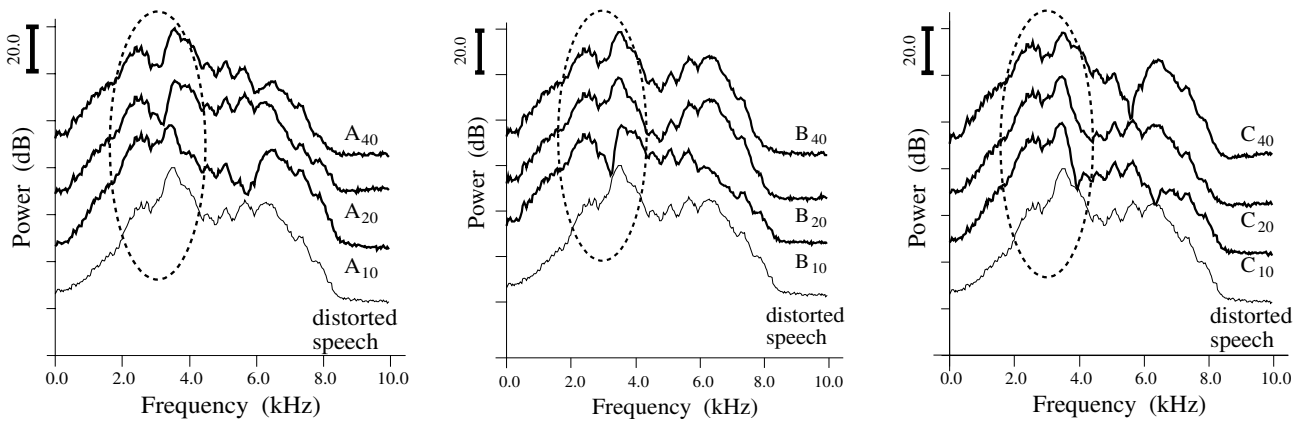


Figure 3: Long term power spectrum of initial distorted speech and speech recovered by operators A, B and C. X_k means speech obtained in k -th generation of interactive GA by operator X. The part inside dotted ellipse shows area around formants' frequencies.

three meters from loudspeakers in a soundproof room. The noise level was less than 30 dB(A). The sound pressure level of presented speech samples were tuned to be 75~80dB at the peak.

3.3 Results of two subject tests

Method of successive categories

This section shows the result of the subjective test of the 13 speeches mentioned in section 3.2 using the method of successive categories. Figure 4 shows the range of each category in Table 1 and the evaluation measure value of the 13 speeches in the distance scale.

What becomes clear from Figure 4 is that: recovered speech samples in 10th and 20th generations of user A are almost of the same quality because they are in a same category. However, one in 40th generation is significantly better than those of 10th and 20th generations in the case of the user A. The same result was obtained in the case of the user C. In the case of the user B, speech in 10th generation is better than those of 20th and 40th generations; the latter two are almost the same.

As a result, it is confirmed that all recovery filters designed by the interactive GA are better than initial distorted speech. This is the most important result; the interactive GA is effective in desining filters that enhance distorted speech. Even if the recovered speech obtained in the minimum generation in our experiment, the 10th generation, the result is better than the initial distorted speech. We think that this fact is useful in reducing the load of users of the interactive GA when it is applied to enhance the distorted speech.

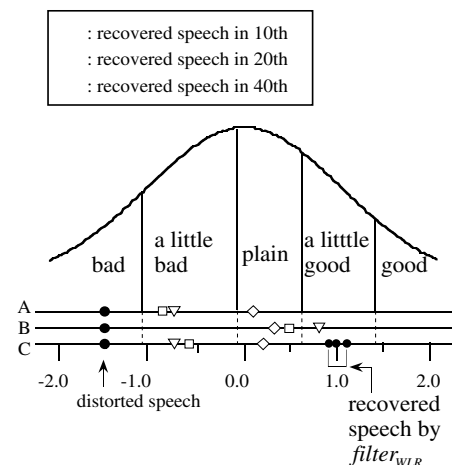


Figure 4: Evaluation of 13 speeches on distance scale obtained by the method of successive categories.

Sheffé's method of paired comparisons

Figure 5 shows distance scales for the three users using Sheffé's method. The average preferences of the initial distorted speech, the speech samples in 10th, 20th, and 40th generations are shown on these scales.

Table 3 shows whether the distance between two average preferences is significant with 1 % degree of statistical risk p . This table shows that the differences between initial distorted speech and all recovered speech designed by operators with interactive GA are significant. This result is same as that of the method of successive categories in the previous section; Sheffé's method also proves that the interactive GA is effective

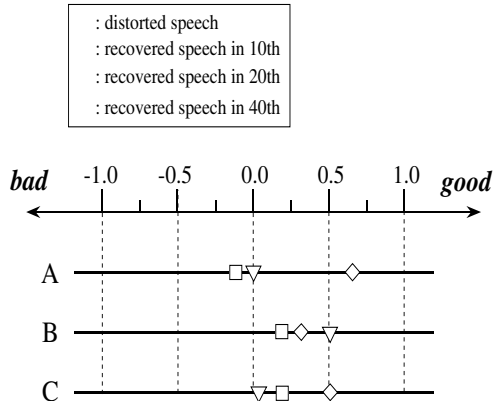


Figure 5: Evaluation of speeches of three operators on distance scale obtained by the Sheffé’s method of paired comparisons

in enhancing distorted speech. It is important to optimize visual or audio systems with respect to human perception. Recent systems, such as codec systems, are designed with considering the human perception, which is more difficult than just considering signal information only. This experimental result implies that interactive GA can be used a good tool to design system that consider human perception.

4 Discussion

Since we have the original clean speech in this experiment, we can define a cost function with the distance scale between recovered speech and original clean speech. Using this cost function corresponds to solving problem by showing the answer. So that, we would get the best recovering filter when the cost function is used. We can evaluate how interactive GA gets near the best condition if we compare the $filter_{human}$ and the recovering filter designed by an optimization method with the cost function.

We design the best filter by using general GA with the cost function as a fitness function. There are many distance scales between two speeches. We use WLR distance. WLR distance is one of spectrum distances measures to compare the difference between two speeches. It is well-known in the auditory research field that human perception is sensitive at around formants’ frequencies. WLR distance measure is defined based on this fact; the error near formats are weighted and emphasized. It has been experimentally reported that WLR distance is more effective than other distance scales in word speech

Table 3: Statistical test with the Sheffé’s method of paired comparisons: D is the initial distorted speech made by a distorting filter; X_k is the speech which is recovered by a recovering filter made by the operator X with k -th generation of interactive GA. \circ means that the difference between two average preferences is significant ($p < 0.01$).

combination	operators		
	A	B	C
D vs. X_{10}	\circ	\circ	\circ
D vs. X_{20}	\circ	\circ	\circ
D vs. X_{40}	\circ	\circ	\circ
X_{10} vs. X_{20}		\circ	
X_{10} vs. X_{40}	\circ	\circ	\circ
X_{20} vs. X_{40}	\circ		\circ

recognition [15].

Let’s call a recovering filters which is designed to minimize the WLR distance between recovered speech and original clean speech using general GA as $filter_{WLR}$.

Figure 6 compares the $filter_{WLR}$ and $filter_{human}$. The frequency characteristics of the $filter_{WLR}$ is more symmetry around 5kHz than that of $filter_{human}$. This means that human operator mainly concentrates on the power of lower frequency which is more important in human perception.

Therefore, the result in section 3 shows that the 32 subjects feel that the closer the recovered speech is to original speech, the more desirable. Calculating WLR distance between recovered speech and original speech, the WLR distance of speech recovered by $filter_{WLR}$ is smaller than that by $filter_{human}$. From the above these points, it is conjectured that $filter_{WLR}$ shows better performance than $filter_{human}$ s. In addition, the number of generation for design of $filter_{WLR}$ is about 600, and is about 15 times of one for $filter_{human}$ s. The WLR distance does not decrease so much regardless of more increase of number of generation. By considering these points, we may conclude that $filter_{WLR}$ in figure 6 is close to the best recovering filter that can be made using 8 order’s FIR filter.

5 Conclusions and future works

This paper has applied interactive GA to enhance the distorted speech. As a result, it has been statistically confirmed that interactive GA is an effective approach to apply to this task. The subjective tests have also indicated that good solutions of this task can be obtained in practical time.

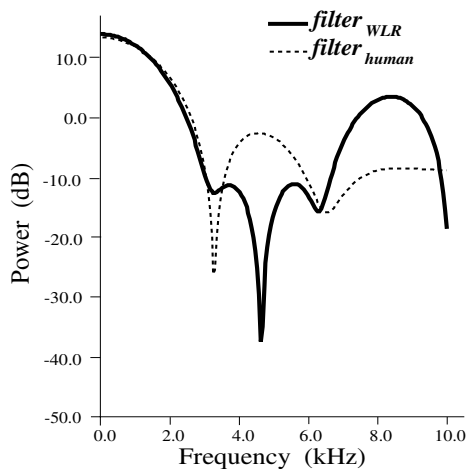


Figure 6: Comparison of two recovering filters. $filter_{WLR}$ is a filter auto-designed to minimize WLR distance using general GA; $filter_{human}$ is a filter designed by human using interactive GA.

This paper has proposed one idea to improve the interface of interactive GA for speech processing. Since the interactive GA cannot show individuals of time signal to a human operator simultaneously and spatially, it is more difficult to apply than to other tasks with no temporally significant behavior. The idea is to combine speech signal with images to assist the user's memory. It is true that there is not enough experimental data to conclude this interface is surely effective, but it is also true that the interactive GA with this visual help succeeded in enhancing speech in this paper.

In future work, it is important to reduce load of a human operator to use interactive GA practically for wide application. In order to reduce the load of the user, many ideas can be imagined: for example, developing a GA with faster convergence to reduce the time that forces human to be a user of interactive GA, improving the interface of presenting individuals, improving the interface of inputting human evaluation.

References

- [1] Holland, J. H.: "Adaptation in Natural and Artificial Systems," University of Michigan Press (1975).
- [2] Goldberg, D. E.: "Genetic Algorithms in Search, Optimization and Machine Learning," Addison-Wesley (1989).
- [3] Caldwell, C. and Johnston, V. S.: "Tracking a Criminal Suspect through "Face-Space" with a Genetic Algorithm," Proc. of 4th Int'l Conf. on Genetic Algorithms (ICGA'91), San Diego, CA, USA, pp.416-421 (July, 1991)
- [4] Smith, J. R.: "Designing Biomorphs with an Interactive Genetic Algorithm," Proc. of 4th Int'l Conf. on Genetic Algorithms (ICGA'91), San Diego, CA, USA, pp.535-538 (July, 1991)
- [5] Sims, K.: "Artificial Evolution for Computer Graphics," Computer Graphics, Vol.25, No.4, pp.319-328 (July, 1991).
- [6] Todd, S. and Latham, W.: "Evolutionary Art and Computers," Academic Press, Harcourt, Brace, Jovanovich (1992).
- [7] Masui, T.: "Graphic Object Layout with Interactive Genetic Algorithms," Proc. of 1992 IEEE Workshop on Visual Languages, Los Alamitos, CA, USA, pp.74-80 (1992).
- [8] Parmee, I. C.: "The Concrete Arch Dam: An Evolutionary Model of the Design Process," Proc. of the Int'l Conf. on Artificial Neural Nets and Genetic Algorithms, Innsbruck, Austria, pp.14-16 (1993).
- [9] Baker, E. and Seltzer, M.: "Evolving Line Drawings," Graphics Interface'94 Proceedings, edited by Wayne A. Davis and Barry Joe. Banff, Alberta, Canada, Morgan Kaufmann Publishers, pp.91-100 (1994).
- [10] Furuta, H., Minami, M. and Watanabe, E.: "A Decision Supporting System for Aesthetic Design of Structure Using Fuzzy Theory and Genetic Algorithms," Proc. of 10th Fuzzy System Symposium, Osaka, Japan, pp.53-56 (June, 1994) (in Japanese).
- [11] Oppenheim, A. V. and Schaffer, R. W.: "Homomorphic Analysis of Speech," *IEEE Trans. on Audio, Electroacoust.*, Vol.AU-16, No.2, pp.221-226 (1968).
- [12] Dixon, W. J. and Massey, F. J.: "Introduction to Statistical Analysis," McGraw-Hill, New York, 342, 343 (1951).
- [13] Sheffé, H.: "An Analysis of Variance for Paired Comparisons," *Am. Statis. Assoc. J.*, Vol.47, pp.381- (1952).
- [14] Ura, S.: "An Analysis of Experiments of Paired Comparisons," *Quality Control*, Vol.16 pp.78-80 (1959) (in Japanese).
- [15] Shikano, K. and Sugiyama, M.: "Evaluation of LPC Spectral Matching Measures for Spoken Word Recognition," *Trans. of the Institute of Electronics and Communication Engineers of Japan, Section E (English)*, Vol.E65, No.5, p.298 (1982).