Proceedings of the 6th WSEAS Int. Conf. on EVOLUTIONARY COMPUTING, Lisbon, Portugal, June 16-18, 2005 (pp80-85)

Text Dependency in Voice Quality Conversion Using Interactive Evolution

YUJI SATO
Faculty of Computer and Information Sciences
Hosei University
3-7-2 Kajino-cho Koganei-shi Tokyo 184-8584
JAPAN
http://www.h3.dion.ne.jp/~y-sato

Abstract: - This paper reports the results of evaluation experiments performed in relation to text dependency in voice quality conversion using interactive evolution. It is important to investigate beforehand whether text dependency exists when considering engineering applications of voice quality conversion technology. In these experiments, for both natural speech recorded with a microphone and synthetic speech generated from text data, prosodic conversion coefficients previously determined for each conversion target by the interactive evolution technique was applied to different text data for the same speaker, and subjects having no knowledge of the purpose of the experiments were asked to evaluate the speech after conversion. We confirmed that prosodic conversion coefficients determined by the interactive evolution technique, while exhibiting speaker dependency, is not text dependent.

Key-Words: - Interactive evolution, Evolutionary computation, Voice quality conversion, Prosodic control

1 Introduction

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 technology [1]. Conventionally, voice conversion confine to the conversion between specified speakers registered beforehand and specialist has to prepare VQ-codebook in advance [2-4]. Therefore, if technology existed that didn't confine to the conversion between specified speakers registered beforehand, and didn't require the work of specialist, and could enable the user to convert speech to a voice of his or her liking, it has a great effect to generate a new market. 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 quality conversion can be felt here. Voice quality conversion technology can also be useful in the world of computer games, especially for multi-player configurations. For example, games that can be operated interactively by voice are to be released for Microsoft's Xbox game console. In these games, 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 quality 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, big markets are now anticipated for a wide range of voice-based applications from the reading out of e-mail and 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.

Against the above background, we have proposed the application of 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 [5]. 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 but also of synthetic speech generated automatically from text data [6]. In this paper, we report on the results of evaluation experiments in relation to text dependency in voice quality conversion using interactive evolution.

Section 2 presents an overview of the voice quality conversion system, section 3 describes techniques for applying evolutionary computation to the problem of voice quality conversion, and section 4 presents the evaluation method and experimental results. The paper concludes with a discussion and summary.

2 Prosodic Coefficient Fitting by Evolutionary Computation

We can consider pitch structure, amplitude structure, temporal structure and spectral structure as the feature quantities for the control of voice quality [7]. On the other hand, it is generally difficult to control dynamic spectral characteristics in real time. Therefore, sought to achieve voice quality conversion by limiting the data to be controlled to pitch data, amplitude data, and temporal structure prosodic data.

2.1 Configuration of the voice modification system

The configuration of our voice modification system is illustrated in Fig. 1. 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 "intelligible," and "childish." 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 rather than at the parameter level, as is done in the usual analysis-synthesis systems.

Figure 2 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

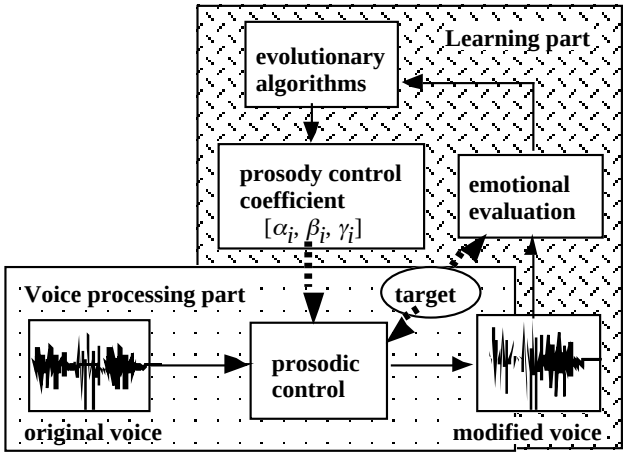


Fig. 1. Block diagram of proposed voice quality conversion system. The system comprises a voice processing part and prosody control coefficient learning part.

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. 3 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) [8] enhancement method. In Fig. 3(a), 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.

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

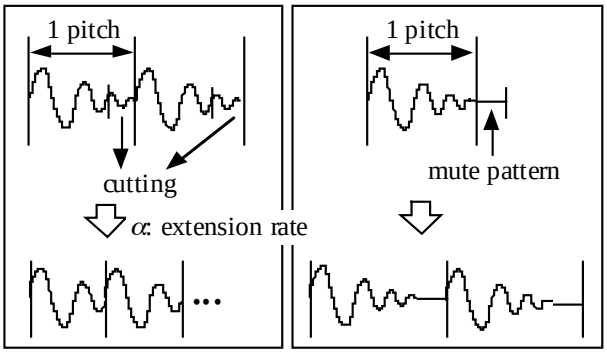
Likewise, in Fig. 3(b), where the contraction rate is denoted as γ, shift *Lc* corresponding to the contraction rate can be expressed by Eq. (2).

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

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.

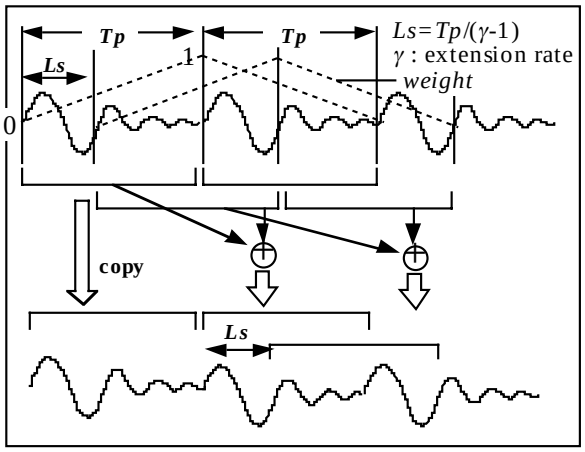
$$\log_{10} W_{i+1}^2 = \log_{10} W_i^2 + \beta \tag{3}$$

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 [9, 10].

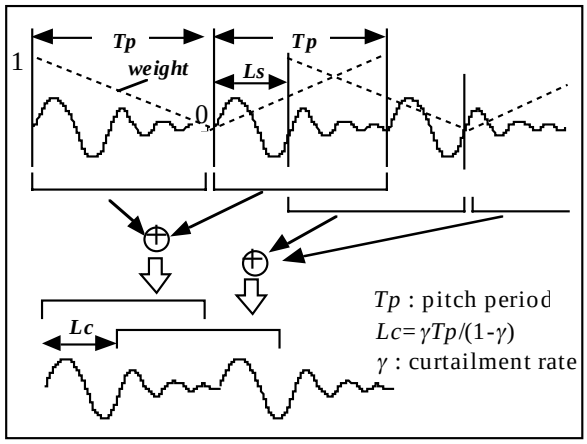


(a) Extension (b) Curtailment

Fig. 2. 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

Fig. 3. 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.

2.2 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:

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

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.

3 An Example of the Application of Evolutionary Computation

We represented the three variables α, β and γ as real numbers, and we defined a chromosome as an array of the form [α, β, γ]. Then, we performed the crossover operation by randomly selecting one variable from among the three array elements and swapping the values of this variable between two parent entities. Figure 4 shows an example of the proposed crossover operation. In this operation, one of the two entities generated by the standard crossover operation is randomly subjected to crossovers in which the average value of each coefficient in the two parent entities [11] 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 to raise the probability that target mutants are in the vicinity of parents and to improve local searching.

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

In the 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 $0.000025I$ [9].

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).

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

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.

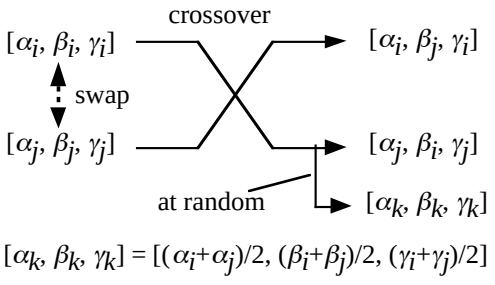


Fig. 4. Example of the genetic manipulation. In this operation, one of the two entities generated by the crossover operation is randomly subjected to crossovers in which the average value of each coefficient in the two parent entities are obtained.

With the number of entities set to 8, the evaluation, selection and genetic manipulation operations were repeated until the input voice data had been converted with an acceptable voice quality. For the evaluation we employed an interactive evolution scheme [12-14] in which humans make subjective (sensory) evaluations of each generation. Based on their evaluation scores, half the entities are discarded, and then the above mentioned crossover and spontaneous mutation operations are performed on parent entities chosen by roulette wheel selection [15], thereby producing child entities to replace the discarded entities.

4 Evaluation Experiments

4.1 Speech stimuli

We performed evaluation experiments using both natural speech recorded with a microphone and synthetic speech generated from text as input data. For natural speech, we asked each of four female speakers to record three different samples of text to generate speech samples S0*i* to S3*i* (*i* = 0 – 2), and then subjected each of these samples to voice conversion with respect to the three subjective categories of "intelligible," "childish," and "masculine" using prosodic conversion coefficients obtained by learning and evolutionary computation to generate speech stimuli S0*ij* to S3*ij* (*i* = 0 – 2, *j* = 0 – 2).

Next, for synthetic speech, we used standard Macintosh software (MacinTalk 3) to generate synthetic speech from three different samples of text each at three levels of voice quality to generate speech samples T0*k* to T2*k* (*k* = 0 – 2), and then subjected each of these samples to voice conversion with respect to the same three subjective categories above in the same manner to generate speech stimuli T0*kl* to T2*kl* (*k* = 0 – 2, *l* = 0 – 2).

4.2 Evaluation method

First, with respect to speech stimulus S00 obtained by having speaker P0 choose and record one sample of text, we determined prosodic conversion coefficients for the three subjective categories of "intelligible," "childish," and "masculine" by performing a search using evolutionary computation as described earlier. We next applied these three prosodic conversion

coefficients to speech stimuli S01 and S02 obtained by having the same speaker record the remaining two text samples. Then, after conversion, subjects were asked to judge whether resulting speech stimuli S01*j* and S02*j* (*j* = 0 – 2) did indeed correspond to the three subjective categories of "intelligible," "childish," and "masculine." This process was applied to all speaker data. There were five subjects in all chosen randomly from men and women in their 20s and 30s having no knowledge of the purpose of these experiments. These five subjects were presented (via speakers) with speech stimuli S*pi* – S*pij* (*p* = 0 – 3, *i* = 0 – 2, *j* = 0 – 2) and T*mkl* – T*mkl* (*m* = 0 – 2, *k* = 0 – 2, *l* = 0 – 2) and asked to evaluate whether voice conversion to the above three subjective categories was successful for these two sets of stimuli. They were asked, in particular, to select one of the following three judgments when making their evaluation: "Close to target represented by subjective category to the same extent as speech for which a prosodic conversion coefficient was determined," "can't say either way," and "not close to target represented by subjective category compared to speech for which a prosodic conversion coefficient was determined." One point was given for "close," zero to "can't say either way," and -1 to "not close." Subjects were allowed to listen to each stimulus several times.

4.3 Evaluation results

Figure 5 shows evaluation results for all subjects with respect to stimuli S*pi* – S*pij* (*p* = 0 – 3, *i* = 0 – 2, *j* = 0 – 2) and T*mkl* – T*mkl* (*m* = 0 – 2, *k* = 0 – 2, *l* = 0 – 2). These results are presented in the form of histograms corresponding to the two judgments "close to target represented by subjective category" and "not close to target represented by subjective category" for each of the three subjective categories of "intelligible," "childish," and "masculine." Here, for 7 speakers' worth of voice quality (from 4 speakers' worth of natural speech and 3 speakers' worth of synthetic speech), these histograms show averages for the evaluations made by five subjects for speech converted by applying previously determined prosodic conversion coefficients to other text samples recorded by the same person. In other words, the prosodic conversion coefficients are applied to the other two text samples for each of the seven speakers so that 14 points represents a perfect score. From the figure, we see that a voice conversion to the target subjective category does indeed take place when

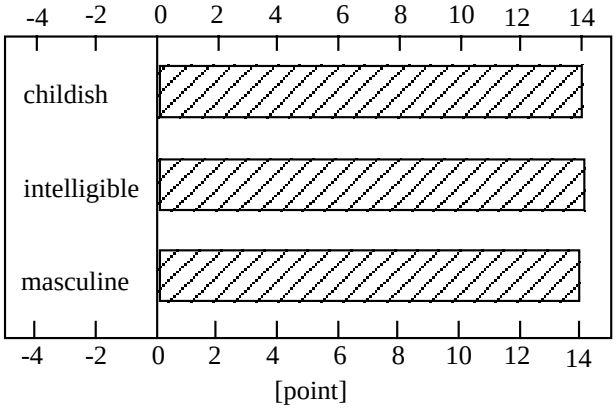


Fig. 5. The results of the judgments of all subjects for voice sample pairs. The results are presented as a histogram for the responses "Close to the target" and " Very unlike the target".

applying a prosodic conversion coefficient obtained by learning using interactive evolution to other text samples for the same speaker. In short, for voice quality conversion using interactive evolution, a prosodic conversion coefficient used for converting speech to a conversion target, while speaker dependent, does not have a text-dependency problem. That is to say, any text dependency that exists is within a range that, for all practical purposes, can be ignored.

5 Discussion

When considering engineering application of voice conversion technology, there are not a few problems to be solved. One of these problems is text dependency. If the assumption is made that a unique, global optimal solution can be found for the prosodic conversion coefficient in the voice quality conversion problem, it is then thought that the problem of text dependency does not exist even if the coefficient is speaker dependent. In actuality, however, in addition to the fact that the speech to be converted is speaker dependent, the evaluation of converted speech also comes to depend on the evaluator and the evaluator's condition at the time of evaluation, for example. As a result, the voice quality conversion problem comes to exhibit time dependency and becomes a problem of finding one of several quasi-optimal solutions within a limited time period. This, in turn, means that the possibility remains of a text-dependency problem depending on the accuracy of the quasi-optimal solution found. If text dependency does exist, it will then be necessary to search out an optimal control

coefficient for each text sample, which would surely limit application of voice quality conversion technology. To investigate this problem, we performed evaluation experiments as described in this paper for both natural speech recorded with a microphone and synthetic speech generated from text data. These computer-based experiments revealed that prosodic conversion coefficients determined by the interactive evolution technique, though speaker dependent, do not have a text-dependency problem, or in other words, that any text dependency that exists is within a range that can be ignored for the most part. This feature is of great benefit when studying the engineering application of voice quality conversion technology such as in computer games and man-personal machine interfaces. Some examples of voice quality conversions performed in this way can be found at http://www.h3.dion.ne.jp/~y-sato/demo/demo1.html.

6 Conclusion

This paper addressed the question of text dependency in prosodic conversion coefficients, an important issue when considering the engineering application of voice quality conversion systems using evolutionary computation as previously proposed. It described, in particular, evaluation experiments performed with respect to both natural speech recorded with a microphone and synthetic speech generated from text data. In these experiments, we applied prosodic conversion coefficients determined beforehand by the interactive evolution technique for certain subjective categories to other text samples recorded by the same speaker, and had subjects who were unfamiliar with the experiments' objective evaluate the speech after conversion. We found that prosodic conversion coefficients determined by the interactive evolution technique, though speaker dependent, are not text dependent.

References:

[1] M. Abe, S. Nakamura, K. Shikano and H. Kuwabara, Voice Conversion through Vector Quantization, in *Proc. of the ICASSP-1988*, 1988, pp. 655-658.

[2] K. Shikano, S. Nakamura, and M. Abe, Speaker Adaptation and Voice Conversion by Codebook Mapping, *IEEE Symposium on Circuits and Systems*, Vol. 1, 1991, pp. 594-597.

[3] E. Moulines and Y. Sagisaka, Voice Conversion: State of Art and Perspectives, *Speech Communication* 16, 1995, pp. 125-126.

[4] M.A. Levent and T. David, Voice Conversion by Codebook Mapping on Line Spectral Frequencies and Excitation Spectrum, in *Proc. of the EuroSpeech*97, 1997

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

[6] Y. Sato, Voice Conversion Using Interactive Evolution of Prosodic Control, in *Proc. of the 2002 Genetic and Evolutionary Computation Conference*, Morgan Kaufmann Publishers, 2002, pp. 1204-1211.

[7] D.H. Klatt and L.C. Klatt, Analysis, Synthesis, and Perception of Voice Quality Variations among Female and Male Talkers, *Journal of Acoustic Society America*, 87(2), 1990

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

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

[10] A. Ghosh and S. Tsutsui (eds.), *Advances in Evolutionary Computing: Theory and Applications*, Springer-Verlag, 2003

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

[12] R. Dawkins, *The Blind Watchmaker*, Long-man, Essex, 1986

[13] K. Sims, Artificial Evolution for Computer Graphics, Computer Graphics, in *Proc. of the ACM SIGGRAPH*, Vol. 25, 1991, pp. 319-328.

[14] H. Takagi, Interactive Evolutionary Computation: Fusion of the Capabilities of EC Optimization and Human Evaluation, in *Tutorial Book of the 2001 Congress on Evolutionary Computation*, IEEE Press, NJ, 2001

[15] D.E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley, MA, 1989