STATISTICAL APPROACH TO VOICE QUALITY CONTROL IN ESOPHAGEAL SPEECH ENHANCEMENT

Kenzo Yamamoto, Tomoki Toda, Hironori Doi, Hiroshi Saruwatari, Kiyohiro Shikano

Graduate school of Information Science, Nara Institute of Science and Technology, JAPAN

tomoki@is.naist.jp, hironori-d@is.naist.jp

ABSTRACT

This paper describes a voice quality control method in statistical esophageal speech enhancement. Esophageal speech is produced by one of the alternative speaking methods for laryngectomees. Its naturalness and intelligibility are much lower than those of natural voices and its voice quality sounds similar even if uttered by different laryngectomees. These issues are alleviated by a statistical voice conversion method from esophageal speech into normal speech (ES-to-Speech) based on eigenvoices. This method is capable of determining converted voice quality using a few target voice samples. In this paper, we propose ES-to-Speech using regression techniques to make it possible to manually control the converted voice quality by manipulating a few intuitively controllable parameters even if no target voice sample is available. The effectiveness of the proposed method is confirmed by experimental evaluations.

Index Terms— Esophageal speech, speech enhancement, voice conversion, voice quality control, kernel regression

1. INTRODUCTION

Laryngectomees who have undergone a total laryngectomy due to an accident or laryngeal cancer cannot produce speech sounds in a usual way because their vocal folds have been removed. They need alternative speaking methods for producing speech sounds. The produced speech is called alaryngeal speech and esophageal speech is a type of alaryngeal speech. In producing esophageal speech, excitation signals are produced by releasing gases from or through the esophagus, and then they are articulated. Esophageal speech sounds more natural than other types of alaryngeal speech, such as electrolaryngeal speech. However, the severe degradation of naturalness and intelligibility of esophageal speech caused by its specific production mechanism is observed compared with normal speech. Moreover, its voice quality is similar even if different laryngectomees speak. Consequently, esophageal speech also suffers from the severe degradation of speaker individuality.

To improve esophageal speech quality, several attempts based on the modifications of acoustic features of esophageal speech using signal processing, such as comb filtering [1] or smoothing of acoustic parameters [2], have been carried out. Although they are useful in the esophageal speech enhancement, quality improvements are still limited since the acoustic features of esophageal speech exhibit quite different properties from those of normal speech. Therefore, it is basically difficult to compensate for those acoustic differences using such a simple modification process.

To significantly improve the naturalness, intelligibility, and speaker individuality of esophageal speech, a statistical method for converting esophageal speech into normal speech (ES-to-Speech) has been proposed [3]. A statistical voice conversion (VC) method [4, 5] is effectively used for converting acoustic features of esophageal speech to those of normal speech. Furthermore, to recover speaker individuality, one-to-many eigenvoice conversion (EVC) [6] has also been implemented for ES-to-Speech to make it possible to flexibly adapt voice quality of the converted speech to that of given target speech samples. Although it is effective for producing more varieties of voice quality, it does not work if any target speech samples with desired voice quality are not available. A technique for manually controlling voice quality is effective to develop a more flexible voice quality control framework.

As one of the statistical parametric speech synthesis techniques capable of manually controlling voice quality of synthetic speech, a multiple regression approach has been proposed in speech synthesis based on hidden Markov model (HMM) [7]. This regression approach has also been applied to one-to-many EVC [8]. In this method, voice quality of various speakers is described by a few voice quality control parameters based on primitive word pairs expressing specific voice quality factors [9]. To manually control converted voice quality without any target voice samples, a subspace spanned by a few representative vectors capturing the specific voice quality factors is formed in a statistical conversion model by extending an eigenvoice-based acoustic modeling technique [10].

Inspired by these methods, we propose voice quality control methods in ES-to-Speech. Manual control of voice quality of the converted speech is achieved by using a multiple regression Gaussian mixture model (MR-GMM) in ES-to-Speech. To further improve the performance of voice quality control in ES-to-Speech, we propose a method for enhancing accuracy of the voice quality control parameters and also propose a more accurate voice quality control method based on kernel regression GMM (KR-GMM). The results of several experimental evaluations are reported to show the effectiveness of the proposed methods.

2. ES-TO-SPEECH BASED ON ONE-TO-MANY EVC

2.1. Training

Using multiple parallel datasets between esophageal speech of a laryngectomee and normal speech of many pre-stored target speakers, the joint probability density function (*p.d.f.*) of a source feature vector of esophageal speech, $\boldsymbol{X}_t$, and a target feature vector of the s^{th} pre-stored target speaker's normal speech, $\boldsymbol{Y}_t^{(s)}$, at frame t is modeled by a one-to-many eigenvoice GMM (EV-GMM) [6] as follows:

$$P\left(\boldsymbol{X}_t, \boldsymbol{Y}_t^{(s)} | \boldsymbol{w}^{(s)}, \boldsymbol{\lambda}\right) = \sum_{m=1}^{M} \alpha_m \mathcal{N}\left(\begin{bmatrix}\boldsymbol{X}_t \\ \boldsymbol{Y}_t^{(s)}\end{bmatrix}; \begin{bmatrix}\boldsymbol{\mu}_m^{(X)} \\ \boldsymbol{\mu}_m^{(Y,s)}\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_m^{(XX)} & \boldsymbol{\Sigma}_m^{(XY)} \\ \boldsymbol{\Sigma}_m^{(YX)} & \boldsymbol{\Sigma}_m^{(YY)}\end{bmatrix}\right), \quad (1)$$

$$\boldsymbol{\mu}_m^{(Y,s)} = \boldsymbol{B}_m^{(Y)} \boldsymbol{w}^{(s)} + \boldsymbol{b}_{m,0}^{(Y)}, \quad (2)$$

This research was supported in part by MIC SCOPE and MEXT Grant-in-Aid for Young Scientists (A).

where $\mathcal{N}(\cdot;\boldsymbol{\mu},\boldsymbol{\Sigma})$ denotes a Gaussian distribution with a mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$. The total number of mixture-components is M. The mixture-component weight α_m, the source mean vector $\boldsymbol{\mu}_m^{(X)}$, and the covariance matrices $\boldsymbol{\Sigma}_m^{(XX)}$, $\boldsymbol{\Sigma}_m^{(XY)}$, $\boldsymbol{\Sigma}_m^{(YX)}$, and $\boldsymbol{\Sigma}_m^{(YY)}$ are tied over every target speaker. On the other hand, the target mean vector of the s^{th} pre-stored target speaker $\boldsymbol{\mu}_m^{(Y,s)}$ is factorized into mixture-component-dependent parameters tied over every target speaker, *i.e.*, a bias vector $\boldsymbol{b}_{m,0}^{(Y)}$ and representative vectors $\boldsymbol{B}_m^{(Y)} = \left[\boldsymbol{b}_{m,1}^{(Y)},\cdots,\boldsymbol{b}_{m,J}^{(Y)}\right]$, and a target-speaker-dependent parameter tied over every mixture-component, *i.e.*, an adaptive vector $\boldsymbol{w}^{(s)} = \left[w_1^{(s)},\cdots,w_J^{(s)}\right]^\top$. Consequently, the EV-GMM has the adaptive vector and a fixed parameter set $\boldsymbol{\lambda}$ consisting of the other parameters.

To convert esophageal speech into normal speech, some parameters of normal speech, such as spectrum, aperiodic components, and F_0, are estimated from a spectral parameter of esophageal speech [3]. For the estimation of spectrum and aperiodic components, we independently train two EV-GMMs modeling the joint *p.d.f.*s of the spectral segment feature of esophageal speech and two features of the normal speech parameters, spectrum and aperiodic components, using corresponding joint feature vector sets. On the other hand, as for the F_0 estimation, a naturally varying F_0 pattern corresponding to perceivable pitch information of esophageal speech is estimated. In other words, a statistical F_0 estimation process for esophageal speech is performed using the VC technique. To achieve such an estimation process, normal speech carefully uttered by a non-laryngectomee so that its pitch sounds similar to that of esophageal speech is newly recorded. Then, its extracted F_0 is used as the target speech parameter to train a standard GMM modeling the joint *p.d.f.* of the spectral segment feature of esophageal speech and the target F_0 feature corresponding to pitch of esophageal speech.

2.2. Adaptation and Conversion

As for the adaptation of spectrum and aperiodic components, each EV-GMM is separately adapted to given target speech samples in an unsupervised manner. An optimum value of the adaptive vector $\boldsymbol{w}$ is determined by maximizing a marginal likelihood $P(\boldsymbol{Y}|\boldsymbol{w},\boldsymbol{\lambda})$ of the EV-GMM for the given target speech features $\boldsymbol{Y}$. For the F_0 adaptation, global mean and standard deviation values are extracted from the given target speech samples.

In conversion, spectrum and aperiodic components are separately estimated from the spectral segment feature of esophageal speech using the corresponding adapted EV-GMMs. F_0 is estimated from the spectral segment feature using the standard GMM. The maximum likelihood estimation method considering dynamic features and the global variance [5] is used in these estimation processes. To adapt global F_0 characteristics to those of the given target speech samples, the estimated F_0 pattern is linearly transformed so as to its mean and standard deviation values over an utterance is equivalent to the target values.

2.3. Limitation

In ES-to-Speech based on one-to-many EVC, only a few arbitrary utterances of the target speech can be used as adaptation data. For instance, even if there were only a small amount of the recorded original voices of laryngectomees before a total laryngectomy, their original voice quality would be recovered. However, those original voices are not always available. Moreover, other voices different from their original ones would be preferred by some people. Therefore, the development of a technique for allowing laryngectomees to customize voice quality as they want is desired.

3. VOICE QUALITY CONTROL IN ES-TO-SPEECH

We propose voice quality control methods in ES-to-Speech to make it possible to manually control converted voice quality. It is essential in voice quality control to design an intuitively controllable parameter to be manipulated. One promising approach is to use perceptual scores expressing specific voice quality factors. In the literature [9], several primitive word pairs to efficiently represent voice quality of various speakers, such as male/female for gender or elder/younger for age, have been extracted through a large-sized perceptual evaluation using normal speech of a lot of speakers. Based on this conventional work, we use perceptual scores on these primitive word pairs as the voice quality control parameter.

3.1. Voice Quality Control Based on MR-GMM in ES-to-Speech

The use of the MR-GMM allows us to manually control converted voice quality [8]. In training, the multiple parallel datasets including esophageal speech and normal speech of many pre-stored target speakers are also used as in the conventional ES-to-Speech based on one-to-many EVC. The perceptual scores on the primitive word pairs are manually assigned to each pre-stored target speaker through listening to his/her natural voices. They are used to form a voice quality control vector of each pre-stored target speaker: *e.g.*, that of the s^{th} pre-stored target speaker is given by $\boldsymbol{w}_c^{(s)} = \left[w_{c,1}^{(s)},\cdots,w_{c,J}^{(s)}\right]^\top$, where individual dimensional components (hereafter voice quality control scores) are given by the manually assigned perceptual scores on J primitive word pairs. Using the multiple parallel datasets and the corresponding voice quality control vectors, the MR-GMM is trained. A model structure of the MR-GMM is the same as that of the EV-GMM but the voice quality control vector $\boldsymbol{w}_c^{(s)}$ is used instead of the adaptive vector $\boldsymbol{w}^{(s)}$. Since the voice quality control vectors of individual pre-stored target speakers are fixed during the training, the resulting model parameters of the MR-GMM are different from those of the EV-GMM. While the representative vectors of the EV-GMM capture dominant voice characteristics over the pre-stored target speakers, those of the MR-GMM capture specific voice quality factors expressed by the primitive word pairs. Moreover, the dimensionality of the voice quality control vector is usually much lower than that of the adaptive vector since the use of a small number of control parameters is preferable in terms of controllability.

Two MR-GMMs for converting the spectral segment feature of esophageal speech into spectrum or aperiodic components of normal speech are independently trained. As for the F_0 control, global mean and standard deviation values of the F_0 pattern are extracted for each pre-stored target speaker, and then, a relationship between those values and the voice quality control vectors over all pre-stored target speakers is modeled by multiple regression analysis.

In the conversion process, the voice quality control vector is manually determined by manipulating each voice quality control score to express the desired voice quality. Then, spectrum and aperiodic components exhibiting the desired voice quality are independently estimated from the spectral segment feature of esophageal speech using the individual MR-GMMs adapted with the determined voice quality control vector. The F_0 pattern is estimated in the same manner as in the conventional ES-to-Speech, and then, it is globally converted so that its mean and standard deviation values are equal to those values estimated from the voice quality control vector.

3.2. Assignment of Voice Quality Control Scores

In the traditional MR-GMM training [8], natural voices of each pre-stored target speaker are used in the assignment of the voice qual-

ity control scores. Therefore, various acoustic features of natural speech, such as local F_0 patterns, duration, and so on, affect the resulting scores more or less even if they are not well controlled in ES-to-Speech. Even in the spectral conversion, some spectral structures affecting voice quality perception would be lost by the effect of statistical generalization. The mismatch between acoustic features affecting the score assignment and those actually controlled in ES-to-Speech causes the performance degradation in voice quality control. To minimize this mismatch, we propose the use of the converted speech in the score assignment. Esophageal speech is converted into each pre-stored target speaker's voice with the target-speaker-dependent GMMs in the conventional ES-to-Speech and the score assignment is performed through listening to the converted speech rather than natural voices. Since the converted speech does not include any varieties of the acoustic features not well modeled by the GMMs, the resulting scores more precisely capture only varieties of the voice quality practically controlled in ES-to-Speech than the scores assigned by listening to natural voices.

3.3. Voice Quality Control Based on Kernel Regression GMM (KR-GMM) in ES-to-Speech

In the MR-GMM, a relationship between each voice quality control vector and the target mean vectors is assumed to be linearly modeled. If such an assumption does not hold, the performance of voice quality control is degraded. To model a more complicated relationship between them, we propose a voice quality control method based on the KR-GMM.

The voice quality control vector is mapped into a high dimensional feature space and linear regression is performed there. In the KR-GMM, given the voice quality control vector $\boldsymbol{w}_c^{(s)}$ of the s^{th} pre-stored target speaker, the d^{th} dimensional component of the target mean vector $\boldsymbol{\mu}_m^{(Y,s)} = \left[\mu_{m,1}^{(Y,s)}, \cdots, \mu_{m,D}^{(Y,s)}\right]^\top$ is modeled by

$$\mu_{m,d}^{(Y,s)} = \sum_{j=1}^{J} \boldsymbol{v}_{j,m,d}^\top \phi\left(w_{c,j}^{(s)}\right), \quad (3)$$

where $\phi(\cdot)$ denotes the function to map each voice quality control score to the feature space. A vector in the feature space $\boldsymbol{v}_{j,m,d}$ can be represented by

$$\boldsymbol{v}_{j,m,d} = \sum_{s=1}^{S} \alpha_{j,m,d,s} \phi\left(w_{c,j}^{(s)}\right), \quad (4)$$

where the number of the pre-stored target speakers is S and the weighting parameter for each data sample is $\alpha_{j,m,d,s}$, which is optimized in kernel regression. Using Eq. (4), Eq. (3) is written as

$$\mu_{m,d}^{(Y,s)} = \boldsymbol{\alpha}_{m,d} \boldsymbol{k}\left(\boldsymbol{w}_c^{(s)}\right), \quad (5)$$

where

$$\boldsymbol{\alpha}_{m,d} = \left[\boldsymbol{\alpha}_{1,m,d}, \cdots, \boldsymbol{\alpha}_{J,m,d}\right], \quad (6)$$

$$\boldsymbol{\alpha}_{j,m,d} = \left[\alpha_{j,m,d,1}, ..., \alpha_{j,m,d,S}\right], \quad (7)$$

$$\boldsymbol{k}(\boldsymbol{w}_c) = \left[\boldsymbol{k}_1^\top\left(w_{c,1}\right), \cdots, \boldsymbol{k}_J^\top\left(w_{c,J}\right)\right]^\top, \quad (8)$$

$$\boldsymbol{k}_j\left(w_{c,j}\right) = \left[k\left(w_{c,j}^{(1)}, w_{c,j}\right), \cdots, k\left(w_{c,j}^{(S)}, w_{c,j}\right)\right]^\top, \quad (9)$$

and $k(\cdot,\cdot)$ denotes the kernel function. In this paper, the radial basis function (RBF) kernel given below is used:

$$k\left(\boldsymbol{x}, \boldsymbol{x'}\right) = \exp\left(-\beta||\boldsymbol{x} - \boldsymbol{x'}||\right), \quad (10)$$

where β is a parameter of the RBF kernel.

As the dependent variables in kernel regression, the target mean vectors of the individual target-speaker-dependent GMMs are used, which are developed by updating only the target mean vectors of the MR-GMM using a parralel dataset corresponding to each pre-stored target speaker. Note that they are no longer represented on the subspace of the MR-GMM since their factorized form is ignored in the update. Let the d^{th} dimensional components of the updated target mean vectors of the m^{th} mixture-component over all pre-stored target speakers be $\hat{\boldsymbol{\mu}}_{m,d}^{(Y,1:S)} = \left[\hat{\mu}_{m,d}^{(Y,1)}, \cdots, \hat{\mu}_{m,d}^{(Y,S)}\right]$. The objective function of kernel regression is given by

$$\epsilon_{m,d}^2 = \left(\hat{\boldsymbol{\mu}}_{m,d}^{(Y,1:S)} - \boldsymbol{\alpha}_{m,d}\boldsymbol{K}^{(1:S)}\right)\left(\hat{\boldsymbol{\mu}}_{m,d}^{(Y,1:S)} - \boldsymbol{\alpha}_{m,d}\boldsymbol{K}^{(1:S)}\right)^\top + \gamma\boldsymbol{\alpha}_{m,d}\boldsymbol{\alpha}_{m,d}^\top, \quad (11)$$

where $\boldsymbol{K}^{(1:S)} = \left[\boldsymbol{k}\left(\boldsymbol{w}_c^{(1)}\right), \cdots, \boldsymbol{k}\left(\boldsymbol{w}_c^{(S)}\right)\right]$ and a parameter for L2 norm regularization is γ. The weighting parameters optimized by the minimization of the objective function are given by

$$\hat{\boldsymbol{\alpha}}_{m,d} = \hat{\boldsymbol{\mu}}_{m,d}^{(Y,1:S)} \boldsymbol{K}^{(1:S)^\top}\left(\boldsymbol{K}^{(1:S)}\boldsymbol{K}^{(1:S)^\top} + \gamma\boldsymbol{I}\right)^{-1}. \quad (12)$$

Consequently, the target mean vectors of the KR-GMM are given by

$$\boldsymbol{\mu}_m^{(Y,s)} = \left[\hat{\boldsymbol{\alpha}}_{m,1}^\top, \cdots, \hat{\boldsymbol{\alpha}}_{m,D}^\top\right]^\top \boldsymbol{k}\left(\boldsymbol{w}_c^{(s)}\right). \quad (13)$$

Two KR-GMMs for the conversion into spectrum and into aperiodic components are separately trained. The kernel regression is also used for the estimation of the F_0 mean and standard deviations for given the voice quality control vector. In this paper, the parameter of kernel function β and the regularization parameter γ are optimized with cross-validation to maximize the conversion performance.

4. EXPERIMENTAL EVALUATIONS

4.1. Experimental Conditions

We recorded 50 phoneme-balanced sentences of esophageal speech uttered by one Japanese male laryngectomee. The same sentences of normal speech uttered by 61 Japanese non-laryngectomees (male 34, female 27) were recorded for training the MR-GMMs/KR-GMMs. Those of pitch-controlled normal speech uttered by one Japanese non-laryngectomee were also recorded for training the standard GMM used for the F_0 estimation. Forty sentences out of the recorded 50 sentences were used for training and the remaining 10 sentences were used for evaluation. Several parameters such as the number of mixture-components of the MR-GMMs/KR-GMMs were experimentally optimized.

The 0^{th} through 24^{th} mel-cepstral coefficients were used as the spectral parameter. As the excitation parameters of normal speech, we used log-scaled F_0 and aperiodic components on five frequency bands (0-1, 1-2, 2-4, 4-6, and 6-8 kHz) for designing mixed excitation. STRAIGHT [11] was used in analysis of normal speech.

To define the voice quality control vector, we used a 5-scaled score (-2: very, -1: somewhat, 0: no preference, 1: somewhat, 2: very) for 5 Japanese primitive word pairs expressing voice quality factors, such as male/female (*gender*), husky/clear (*clearness*), elder/younger (*age*), deep/thin (*deepness*), and weak/strong (*forcefulness*). One Japanese male subject assigned these scores to each of the pre-stored target speakers. The assigned scores for each word pair were normalized into Z-score (zero mean and unit variance) over all pre-stored target speakers. Although the results can not be shown here due to lack of space, we confirmed that the use of the scores assigned to the converted speech as described in **Section 3.2** makes results of voice quality control more stable compared with those when using the scores assigned to natural voices.

4.2. Experimental Results

4.2.1. Evaluation of Controllability

We conducted subjective evaluations of voice quality control. The number of listeners was 10. Five test sentences were used. For each sentence, we synthesized 5 samples of the converted speech by varying only one voice quality control score from -2 to 2 in 5 steps while setting the other voice quality control scores to zero. The converted speech samples when setting every voice quality control score to zero were also synthesized as reference speech. Each listener compared the voice quality of the converted speech with that of the reference speech using a 5-scaled score (-2: very, -1: somewhat, 0: no difference, 1: somewhat, 2: very) for the primitive word pair corresponding to that on the varied voice quality control score.

Figure 1 shows a result of voice quality control on *deepness*. The perceptual score well correlates to the setting of the voice quality control score in both the EV-GMM and the KR-GMM. We have found that good correlation between those two scores can also be observed when manipulating other voice quality control scores. These results show that the proposed methods are capable of effectively controlling voice quality of the converted speech in ES-to-Speech by manipulating the voice quality control vector.

We can also see that the KR-GMM is capable of generating converted speech of which voice quality varies more widely and is closer to the setting of voice quality control score compared with the MR-GMM. Discussion on a comparison between the MR-GMM and the KR-GMM is shown below.

4.2.2. Evaluation of Naturalness

We also conducted opinion tests of naturalness. Naturalness of the converted speech was evaluated using a 5-scaled opinion score (from 1: very bad to 5: excellent). The other experimental conditions were the same as those in the previous evaluations on controllability.

Figure 2 shows a result when manipulating only the voice quality control score on *deepness*. We can see that naturalness of the converted speech starts to be degraded when the voice quality control score is set to too large or too small values. Therefore, it is better to keep the voice quality control score in a reasonable score range. Although there is no significant difference between the MR-GMM and the KR-GMM in this figure, the KR-GMM is capable of controlling target voice quality more widely than the MR-GMM using the same range of voice quality control score as shown in **Figure 1**. We have also found that the KR-GMM can significantly reduce the degradation of naturalness caused by setting the voice quality control score to too small values compared with the MR-GMM when manipulating another voice quality control score. These results suggest that the KR-GMM is more effective than the MR-GMM for voice quality control in ES-to-Speech.

5. CONCLUSION

This paper has described novel methods to intuitively control voice quality in esophageal speech enhancement based on statistical voice conversion. The multiple regression Gaussian mixture model (MR-GMM) has been implemented for a statistical conversion process from the esophageal speech into normal speech (ES-to-Speech). Moreover, the kernel regression GMM (KR-GMM) has been proposed to further improve the controllability of the voice quality. The experimental results have showed that 1) our proposed methods allow a laryngectomee to control voice quality of the converted speech by manipulating a few control parameters on primitive word pairs expressing specific voice quality factors and 2) the KR-GMM yields better performance in voice quality control than the MR-GMM.

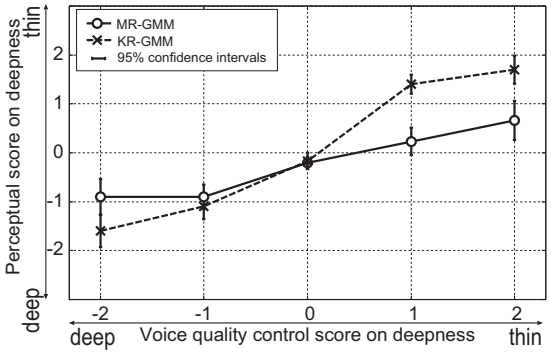

Fig. 1. Result of evaluation of voice quality controllability.

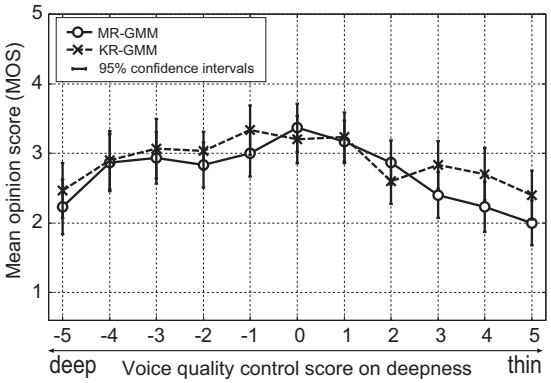

Fig. 2. Result of evaluation of naturalness.

6. REFERENCES

[1] A. Hisada and H. Sawada. Real-time clarification of esophageal speech using a comb filter. *Proc. 4th Intl Conf. Disability, Virtual Reality & Assoc. Tech.*, pp. 39–46, Veszprem, Hungary, 2002.

[2] K. Matsui, N. Hara, N. Kobayashi, and H. Hirose. Enhancement of esophageal speech using formant synthesis. *Acoust. Sci. & Tech.*, Vol. 23, No. 2, pp. 69–76, 2002.

[3] H. Doi, K. Nakamura, T. Toda, H. Saruwatari, and K. Shikano. Esophageal speech enhancement based on statistical voice conversion with Gaussian mixture models. *IEICE Trans. Inf. & Syst.*, Vol. E93-D, No. 9, pp. 2472–2482, 2010.

[4] Y. Stylianou, O. Cappe, and E. Moulines. Continuous probabilistic transform for voice conversion. *IEEE Trans. SAP*, Vol. 6, No. 2, pp. 131–142, 1998.

[5] T. Toda, A.W. Black, and K. Tokuda. Voice conversion based on maximum likelihood estimation of spectral parameter trajectory. *IEEE Trans. ASLP*, Vol. 15, No. 8, pp. 2222–2235, 2007.

[6] T. Toda, Y. Ohtani, and K. Shikano. One-to-many and many-to-one voice conversion based on eigenvoices. *Proc. ICASSP*, pp. 1249–1252, Hawaii, USA, Apr. 2007.

[7] T. Nose, J. Yamagishi, T. Masuko, and T. Kobayashi. A style control technique for HMM-based expressive speech synthesis. *IEICE Trans. Inf. & Syst.*, Vol. E90-D, No. 9, pp. 1406–1413, 2007.

[8] K. Ohta, Y. Ohtani, T. Toda, H. Saruwatari, and K. Shikano. Regression approaches to voice quality control based on one-to-many eigenvoice conversion. *6th ISCA Speech Synthesis Workshop (SSW6)*, pp. 101–106, Bonn, Germany, Aug. 2007.

[9] H. Kido and H. Kasuya. Everyday expressions associated with voice quality of normal utterance —Extraction by perceptual evaluation—. *J. Acoust. Soc. Jpn.*, Vol. 57, No. 5, pp. 337-344, 2001 [in Japanese].

[10] R. Kuhn, J. Junqua, P. Nguyen, and N. Niedzielski. Rapid speaker adaptation in eigenvoice space. *IEEE Trans. SAP*, Vol. 8, No. 6, pp. 695–707, 2000.

[11] H. Kawahara, I. Masuda-Katsuse, and A.de Cheveigné. Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F_0 extraction: Possible role of a repetitive structure in sounds. *Speech Communication*, Vol. 27, No. 3-4, pp. 187–207, 1999.